Video Traffic: Characterization, Modelling and Transmission

C. H. Liew

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Communication Systems Research
School of Electronics and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

February 2006

Summary

Ubiquitous wireless multimedia communications are becoming a reality. Users will be able to communicate and access information etc. just about anytime, anywhere and will always be connected to the network. Telecommunication operators now consider multimedia services as revenue generators. However, the success of multimedia services in attracting customers also depend on the capability of the network to guarantee Quality of Service (QoS). Hence, the topic of multimedia QoS provisioning constitutes an important research area.

The research of multimedia QoS provisioning can be divided into two parts. The first part includes the study of application QoS requirements, traffic characterization, and mathematical modelling of the traffic. A traffic model is important for evaluation of communication system performances by either analytical techniques or software simulations without costly hardware prototypes. The second part includes the design and development of multimedia resource allocation and scheduling algorithms in packet switched mobile networks. Resource allocation ensures that sufficient amounts of resources are allotted to users whereas scheduling ensures that user access to the allotted resources is given on a timely basis.

Thus the objectives of this thesis are to characterize and model multimedia traffic as well as to propose efficient resource allocation and scheduling algorithms. In the first part, two new accurate Variable Bit Rate (VBR) video traffic models are presented together with their performance evaluations. Leading on from there, enhancements are added so that the model becomes reconfigurable to encoder parameters. The second part of the thesis deals with multimedia resource allocation and scheduling. A joint channel- and user- aware multiple user resource allocation algorithm for multimedia traffic is studied. Finally, the design of a low complexity scheduling algorithm for delay sensitive multimedia traffic in wireless network is proposed.

Keywords: Traffic Modelling, Multimedia Communication, Resource Allocation, Scheduling

Email : c.liew@surrey.ac.uk
WWW : http://www.ee.surrey.ac.uk/CCSR

Acknowledgements

I would like to express my gratitude to Professor A. M. Kondoz and Dr. C. K. Kodikara for their invaluable advices and time on my PhD work. I would also like to thank Multimedia Group of Centre for Communication Systems Research (CCSR) for funding my study. Not to forget about my colleagues in the centre, their friendship and support have made my stay enjoyable. Special thanks to secretarial support offerred by Anne Rubin and Fiona Wilson. Finally but importantly, the support, love, and encouragement of my parents and sister are highly appreciated.

Contents

Chapter 1

Introduction

1.1 Project Motivations and Objectives

Ubiquitous multimedia communications are becoming a reality as mobile operators around the world are deploying the third generation Universal Mobile Telecommunication Service (UMTS) [1] and fourth generation Worldwide Interoperability for Microwave Access (WIMAX) [2]. While mobile operators expect to generate revenue from multimedia services, the success of multimedia services in attracting customers still depends on the capability of the network to guarantee Quality of Service (QoS). Thus, multimedia QoS provisioning constitutes an important research area which can be divided into two parts. The first part includes the study of application QoS requirements, traffic characterization, and mathematical modelling of the traffic characteristics. Upon understanding the traffic characteristics, efficient resource management schemes can be designed [3]. The second part of multimedia QoS provisioning is the resource management scheme, which consists of resource allocation and scheduling.

Traffic characterization and modelling facilitates the design and optimization of

communication systems [4]. Traffic modelling is important as it is used to evaluate the performance of a communication system by analytical techniques or by software simulation. This avoids the need to build costly hardware prototypes. For analytical techniques, Markov queuing theory is commonly used, e.g. [5] [6] [7], to derive a range of properties such as buffer overflow probability, average delay and delay probability distribution. These properties give a performance indication to the system under study. However, the theoretical tractability of the system performances is becoming increasingly difficult as system complexity grows. For to this reason, software simulation is generally favored over analytical techniques. In this context, traffic models are employed as synthetic traffic generators to drive the simulation. Various system performances can then be estimated by simulation. Although real traffic traces can be used for simulation, this requires a huge amount of storage. In contrast, the traffic model is compact, requires less storage space and is able to instantiate different traffic traces with the same statistical behaviour.

Resource management is an important multimedia QoS provisioning topic that ensures the users requested QoS is satisfied, while efficiently utilizing scarce radio resources. Two important resource management schemes are resource allocation and scheduling techniques. The resource allocation techniques play an important role in allotting sufficient amounts of system resources to users to sustain user-requested QoS, while at the same time maximizing resource utilization and revenue. As for scheduling, it allows the user to access allotted system resources on a timely basis. Timely access to the allotted system resources reduces the transmission delay, ensuring rapid system response is perceived at the application. The design of resource allocation and scheduling algorithms is not a simple task as it involves several contradicting dimensions e.g. QoS requirement and efficient resource utilization. Also the research in this area are ongoing and algo-

rithms should be optimized for a particular air interface (e.g. CDMA or OFDM).

Following on from previous discussions, thus the objectives of this thesis is set out to characterize and model multimedia traffic. Although multimedia traffic consists of a combination of video, voice, audio and text, this thesis concentrates on video traffic due to its high bandwidth consumption nature and stringent QoS requirements. Due to the reasons discussed, the thesis concentrates on traffic modelling for synthetic traffic generation. The study concentrates on MPEG4 encoded video traffic since it is widely used and a chosen codec for Universal Mobile Telecommunication System (UMTS) [8]. The second objective of the project is the investigation and evaluation of multimedia resource allocation and scheduling algorithms with the proposed MPEG4 video traffic model. In particular, a resource allocation algorithm that jointly considers channel conditions and user level quality is proposed. Finally, a low complexity delay sensitive resource scheduling algorithm is proposed and evaluated using the MPEG4 video traffic model.

1.2 Research Achievements

The research achievements resulting from the work described in this thesis can be summarized below

- Detailed study of the MPEG4 video traffic characteristics.
- Two new and accurate MPEG4 Variable Bit Rate (VBR) video traffic models called Spatial-Renewal-Process-MM (SRP-MM) and Nested-AutoRegressive-MM (NestedAR-MM) are proposed based on a MultinoMial (MM) technique. [Appendix A,1][Appendix A,3]

- A Generalized Video Traffic Model (GVTM) is proposed. GVTM extends SRP-MM to allow the simulation of VBR traffic with different quantization parameters in real time, avoiding time consuming model parameter re-estimation. [Appendix A,2][Appendix A,4]
- The design of a joint channel- and user- aware multiple user resource allocation algorithm for video traffic in wireless networks. [Appendix A,5]
- The design of a low complexity scheduling algorithm for delay sensitive multimedia traffic in wireless network. [Appendix A,5]

A number of publications have been produced as a result of the conducted research conducted. The journal and conference publications written by the author are listed in Appendix A. The research work has contributed to the European IST NEWCOM project.

1.3 Structure of Thesis

Chapter 1 gives the motivations and objectives of the project work. Chapter 7 summarizes the research work conducted and its achievements. It also outlines some of the potential research area as a continuation of this work. The remaining chapters are summarized below.

1.3.1 Contents of Chapter 2

Chapter 2 outlines the background to the project. The chapter starts with the fundamental concepts of video traffic modelling. Then the classical tools for video traffic modelling are described, including Markov, Regression and Long Range Dependent (LRD) mathematical models. These tools can be used independently

or in combination to model the empirical traffic characteristics. The final part of the chapter discusses the topic of multimedia resource allocation and scheduling in wireless environments. The topics covered include radio resource management, degree of freedom in system parameters, system optimization objectives and cross layer optimization.

1.3.2 Contents of Chapter 3

Chapter 3 features two video traffic modelling techniques for MPEG4 encoded video based on the detail examination of video traffic characteristics. The first model combines the Spatial Renewal Process with MultinoMial (MM) technique to model video traffic. The model is called (SRP-MM). The second model combines the Nested AutoRegressive process with MM (Nested-AR-MM) to model video traffic. The proposed SRP-MM and Nested-AR-MM are validated by simulation in terms of marginal distribution matching, autocorrelation matching and packet loss rate matching.

1.3.3 Contents of Chapter 4

SRP-MM and Nested-AR-MM captures video traffic characteristics accurately for a pre-defined encoder parameter set, but require time consuming model parameter re-estimation process if other encoder parameter sets are desired. Chapter 4 describes a Generalized Video Traffic Model (GVTM) that overcomes this difficulty by using an adaptive frame size model. GVTM is re-configurable for the generation of video traffic with different quantization parameter set in real time. This also allows the simulation of adaptive source rate video codec. GVTM is evaluated by using the standard validation techniques, i.e. marginal distribution matching, autocorrelation matching and packet loss rate matching. Its prediction accuracy is shown for different quantization parameter sets.

1.3.4 Contents of Chapter 5

In order to study multimedia resource allocation and scheduling schemes in wireless environments, a link level simulator and a system level simulator for the Orthogonal Frequency Division Multiplexing (OFDM) air interface are implemented. OFDM is considered as the wireless air interface due to its robustness against frequency selective fading and is suitable for high rate multimedia transmission. In Chapter 5, the fundamentals of OFDM and its advantages are reviewed first. Then the implementation of a link level simulator is then described. Finally, a system level simulator is implemented. The system level simulator utilizes the block error rate performance curves from link level simulation to imitate the wireless channels behaviour. The Effective Exponential SIR Mapping (EESM) technique is used as an interface between the system level simulator and the link level simulator.

1.3.5 Contents of Chapter 6

Chapter 6 studies the problem of multimedia resource allocation and scheduling for multiple users. First, a system model for the study is introduced. Secondly, a user- and channel- aware optimized Genetic Algorithm Based Resource Allocation (GABRA) algorithm is proposed and evaluated. Finally, a low complexity scheduling algorithm called Minimum Queue Length Ratio Scheduler (MQLRS) is proposed for delay sensitive multimedia traffic. The characteristics and advantages of MQLRS are demonstrated by simulation. GABRA and MQLRS are studied for the Frequency Hopped Orthogonal Frequency Multiple Access (FH-OFDMA) system due to its robustness to multipath interference and intercell interference.

Chapter 2

Background

Chapter 2 is divided into three sections. The first section describes the fundamental concepts in video traffic modelling. These concepts are also used in traffic modelling of other applications such as ethernet, FTP, web application and etc. In the second section, classical mathematical tools for video traffic modelling are described. The third section gives an overview of fundamental issues of multimedia resource allocation and scheduling in wireless environments.

2.1 Fundamentals of Video Traffic Modelling

This section explains the fundamental concepts of video traffic modelling. They include heavy-tailed distribution, Long Range Dependency (LRD) and fractals. Before the discussions, it is necessary to introduce the variable X_k

$$X_k = N[kT_s] - N[(k-1)T_s], \tag{2.1}$$

where $N[T]$ is the counting process from time $t = 0$ to $t = T$. Here X_k may represent the number of packets, cells or bytes that arrived during the k^{th} time interval of duration T_s. X_k will be used as defined unless otherwise specified.

2.1.1 Heavy-tailed Distribution

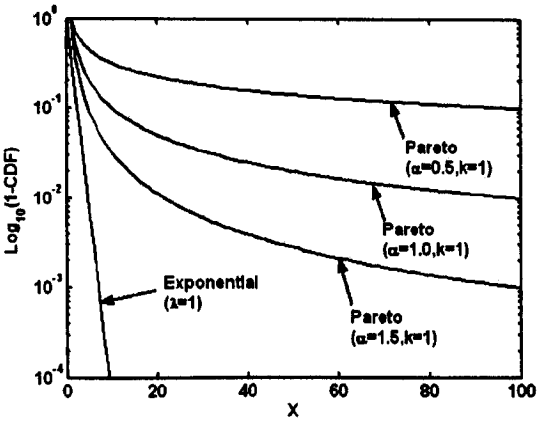


Figure 2.1: Pareto distribution

A random variable U is considered heavy-tailed if its Complementary CDF (CCDF) satisfies the following

$$P\{U \geq x\} = 1 - F(x) \sim c_1 x^{-\alpha} \quad \text{for} \quad x \to \infty, \quad 0 < \alpha < 2, \tag{2.2}$$

where $F(x)$ is the CDF and c_1 is a positive finite constant. As an example, the simplest heavy tailed distribution is the Pareto distribution

$$F(x) = 1 - \left(\frac{k}{x}\right)^{\alpha}, \quad x > k, \quad \alpha > 0. \tag{2.3}$$

The Pareto CCDF is plotted in Fig. 2.1. It can be seen that the tail of Pareto distribution decreases at a slower rate than the exponential distribution; hence the term heavy-tailed. The heavy-tailed property is commonly observed in the traffic modelling literature. For instance, the marginal distribution of data and VBR video traffic arrival process X_k were observed to exhibit the heavy-tailed property [9] [10]. The data transmission time was also observed to be heavy-tailed [11]. Although in practice the probability of the tail (large values) may be very small, they need to be considered during the modelling process as opposed

to considering it as a statistical outlier. This is because the network experiences packet loss due to such large values (large burst of traffic). If they are ignored, empirical queuing performance may be underestimated if the traffic model is used as a traffic generator in the simulation.

2.1.2 Long Range Dependency

One of the widely known characteristic in VBR video and data traffic is the so-called Long Range Dependency (LRD) [9] [10]. LRD is characterized by slow decaying autocorrelation function (ACF) of X_k. That is to say that the ACF of X_k does not die out after a large lag. To be more precise, the ACF $\rho(k)$ has the following form [12]

$$\rho(k) \sim c_3 k^{-\beta}, \quad as \quad k \to \infty, \quad 0 < \beta < 1, \tag{2.4}$$

where c_3 is a finite positive constant. This results in non-summable ACF, i.e. $\sum_{k=0}^{\infty} \rho(k) = \infty$. LRD is an indication of strong temporal correlation over a large lag. This property manifests itself as an occurrence of a pronounced cluster of consecutive large or consecutive small values.

LRD is often explained using an ON-OFF source model [13] as shown in Fig. 2.2. For example, a user might be actively browsing a website thus requiring data transfer (ON transmission time), and after the site is loaded, the user will spend some time reading and no information is downloaded during that moment (OFF transmission time). Willinger [13] has found that ON and OFF transmission time

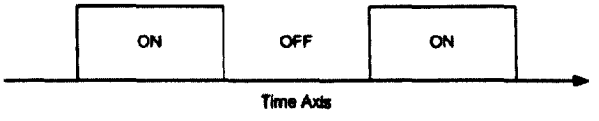


Figure 2.2: ON-OFF source traffic model

tends to be heavy-tailed and analytically proved that the aggregation of many flows with heavy-tailed transmission time causes LRD.

LRD type traffic represents a long burst of high traffic (or low traffic) arrival during a high activity period which can hardly be absorbed by merely increasing the buffer size [14]. LRD can have a detrimental effect on network performance as suggested in [14] [15]. Grossglauser [14] have found that increasing the buffer size has little impact on decreasing the packet loss rate for long range dependent data traffic. Erramilli *et al.* [15] have shown that the buffer overflow probability of LRD traffic was several orders higher than the traditional Markov-based modelling approach that have short range dependent behavior. This is because the Markov based technique would not be able to capture the autocorrelation behavior long enough to reflect the inherent burstiness in the traffic.

2.1.3 Fractals

Fractals were introduced by Mandelbrot. They are geometric objects that exhibit a highly irregular appearance. The mathematical properties of fractals appear to be a valuable tool to analyze scale-invariant behaviour of network traffic. The discovery of fractal-like behaviour in network traffic was a major breakthrough and revolutionized traffic modelling research [12]. Fractals can be divided into two types: Monofractal and Multifractal. Monofractal and multifractal based techniques were used in the past to analyze video traffic [9] [17].

Monofractals are also called self-similar. A self-similar object is an object where part of it looks very similar to itself as a whole. For example, Fig. 2.3(a) shows a self-similar process while Fig. 2.3(b) shows a non- self-similar process. It can be seen that the zoomed in version of self-similar process looks similar to the

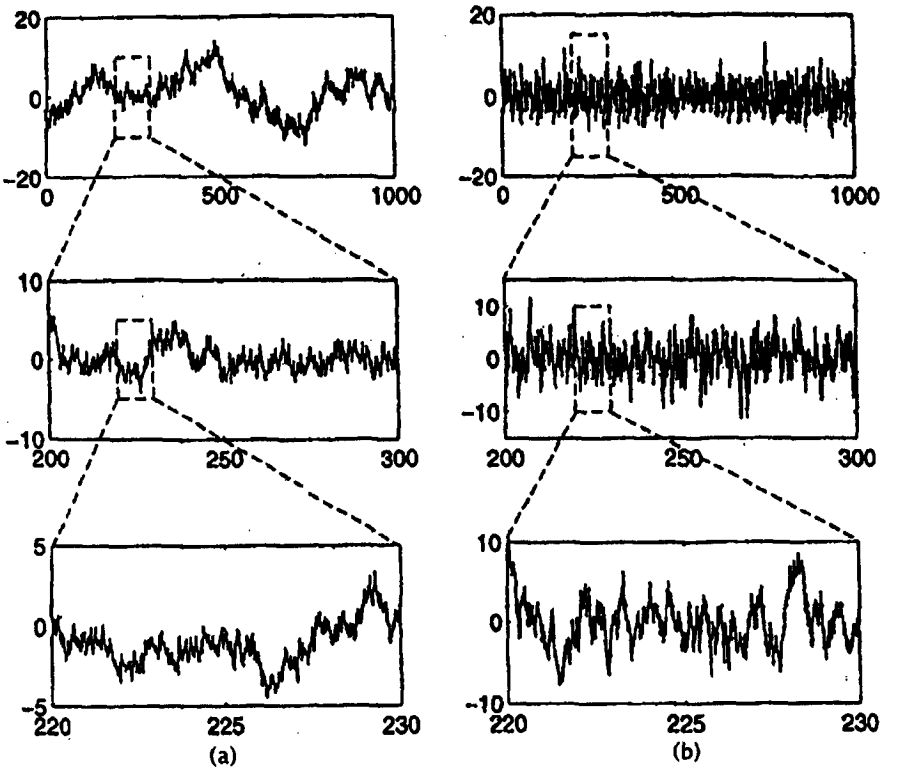


Figure 2.3: Self-similar and non- self-similar processes [16]

original one. A more rigorous mathematical definition of self-similar is as follows: a process X_k is called exactly self-similar (or Monofractal) with scaling exponent (or Hurst parameter) $H = 1 - \frac{\beta}{2}$ if for all $m = 1, 2, ...$

$$X_k = m^{1-H} X_k^{(m)}, \tag{2.5}$$

where the equality is in the distributional sense. In other words, when the aggregated process $X_k^{(m)}$ is scaled with factor m^{1-H}, it has the same statistical properties to X_k. Follow from equation Eq. (2.5)

$$\rho^{(m)}(\tau) = \rho(\tau), \quad \tau > 0. \tag{2.6}$$

A weaker condition for self-similarity is the following: A process X_k is called

asymptotically self-similar if

$$\rho^{(m)}(k) \to \rho(k), \quad m \to \infty. \tag{2.7}$$

Mathematically, self-similarity manifests itself in the following equivalent ways: (i) self-similar processes are long-range dependent; (ii) self-similar processes have slow decaying variance.

While self-similarity relates to traffic characteristics over a large time scales, it was observed that video traffic [17] and TCP/IP traffic [18] exhibits multifractal behaviour over a small time scales. The multifractal [19] generalizes the monofractal and exhibits richer behaviour. Multifractal is studied via its increment process of the cumulative process $N(x)$. Here $N(x)$ is defined as the amount of traffic arrived from time $t = 0$ to $t = x$. At intervals of $[x_0 + \delta x]$, the arrived traffic (or the increment process) is $N(x_0 + \delta x) - N(x_0)$. The traffic is said to exhibit multifractal behaviour with scaling exponent $\alpha(x_0)$ if the increment process behaves like $(\delta x)^{\alpha(x_0)}$ as $\delta x \to 0$. Informally, the traffic with the same scaling exponent at all instants x_0 is called monofractal or self-similar, for which $\alpha(x_0) = H$, while traffic with non-constant scaling exponent $\alpha(x_0)$ is called multifractal. It was argued that various complex functions and responses of protocol layers contribute to such small time scale behaviour. It has been shown that multifractal captures the statistical behavior more accurately than monofractal model [20].

2.2 Mathematical Modelling Tools

This section describes the classical video traffic modelling tools that are widely used. They include Markov, Regression and LRD type tools [21] [22]. These tools have also been used in the literature for the modelling of traffic for various applications e.g. FTP and web.

2.2.1 Markov Model

In Markov type models, the traffic bit rate is quantized into a number of discrete states $S = \{s_1, s_2, ..., s_M\}$. Traffic bit rate variation is modelled as a transition between states. Each transition path is associated with a certain probability indicating the frequency of changes from one particular state to another. An example of a 3-state Markov process is shown in Fig. 2.4. Let X_k be a random variable defining the state at time k, then the series $\{X_k\}$ is called the discrete Markov chain if the probability of being in the next state, $X_{k+1} = s_{j+1}$, only depends on the current state $X_k = s_j$. This is known as the Markov property [23].

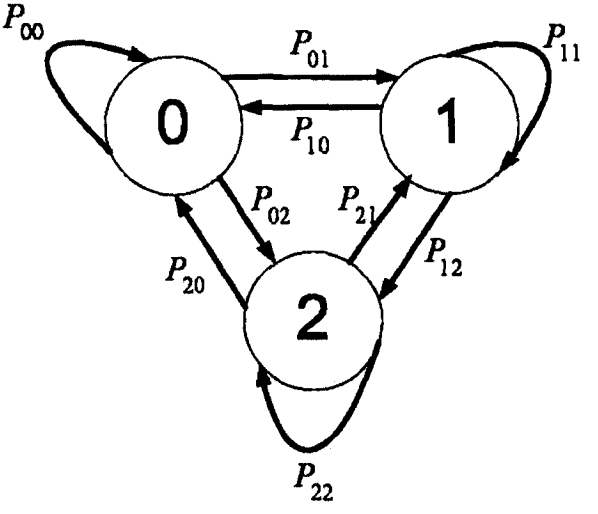


Figure 2.4: A simple markov chain

One of the important classes of the Markov model for traffic modelling is the Markov Modulated Process (MMP) [21] [23]. In MMP, a stochastic process is used to modulate the parameter of another stochastic process. Markov Modulated Fluid Process (MMFP) [23] is an example of MMP. The Markov chain was used to modulate the arrival rate, λ_i. λ_i is state dependent and stays constant during its sojourn time in a markov state. The MMFP was used to model video telephony

traffic in [24]. A more elaborate MMP based model in [25] is described as follows. The MPEG Group of Picture (GOP) process was quantized into four states, S, M, L and XL, and the probability of transition between states are estimated. S, M, L and XL correspond to states from low arrival rate to high arrival rate. For each state, the marginal distribution of bit rate was estimated. In order to generate the MPEG GOP size, the current state of Markov chain was determined. Given the state, the GOP size was generated randomly from the marginal distribution of the bit rate for that particular state. The state sojourn time distribution was defined by a heavy-tailed distribution. Another similar MMP model was proposed in [26]. The model was defined by two markov chain processes where one Markov process is nested insider another Markov process. The outter markov chain considers the visual scene state changes, while the inner state models the GOP bit rate process within each scene state.

2.2.2 Regression Model

The regression model is another common tool in traffic modelling. Basically it is assumed that the current values can be predicted from the weighted sum of past values plus some random noise. The Regression model has the following form [23]

$$X_k = \sum_{r=1}^{p} \alpha_r X_{k-r} + \sum_{r=1}^{q} \beta_r \epsilon_{k-r} + \epsilon_k, \quad k > 0, \tag{2.8}$$

where a_0, α_r and β_r are constants and ϵ_k are zero-mean, Independent Identically Distributed (IID) normal random variables. The model in Eq. (2.8) is called Autoregressive Moving Average (ARMA(p,q)) [27]. It is also called AutoRegressive (AR(p)) if $\beta_r = 0$, or Moving Average (MA(q)) if $\alpha_r = 0$. AR(p) is the most commonly used regression model for video traffic modelling. For example, Krunz [28] have used the AR(p) process to model the I frame of MPEG streams. In [29], Golaup have used AR(p) to model MPEG4 video at GOP level. Despite the popularity of AR(p), it is inherently short range dependent (SRD) due to

their fast decaying ACF. This contradicts the observations of LRD in video traffic [30] and may not be suitable for all types of video. An extension called Nested AutoRegressive(Nested-AR) [31] attempts to overcome the weakness of AR(p) for long range dependent video traffic modelling. Note that X_k is normal distributed because ϵ_k is normal distributed. X_k need to be transformed to the desired marginal distribution by probability integral transform (see Appendix B.2).

2.2.3 LRD Model

Wavelets

Wavelets is a class of powerful tools suitable for traffic modelling. When wavelet decomposition is performed on network traffic with LRD features, the resulting wavelet coefficients are less correlated and easier to model. Wavelet traffic models have shown consistently better performance in terms of their closeness of queuing performance to the empirical trace [32]. The plausible wavelet decorrelation property [32] simplifies the network traffic modelling process has make wavelet a popular tools for modelling.

In short, a wavelet is a set of orthonormal bases which can be used to represent a signal as function of time. A discrete wavelet has the following form

$$\phi_j^m(t) = 2^{-j/2}\phi(2^{-j}t - m), \tag{2.9}$$

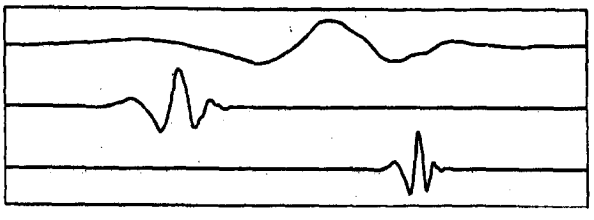


Figure 2.5: Daubechies wavelet

where m represents the time translation while j represents the scaling factor. By varying the scaling factor j, the wavelet is dilated or compressed. Fig. 2.5 shows an example of Daubechies wavelets. The upper two plots of wavelets are the dilated and shifted version of the original wavelet at the lowest plot. Using wavelet such as the one shown in Fig. 2.5, a signal $X(t)$ can be decomposed by

$$d_j^m = \sum_{t=0}^{2^K-1} X(t)\phi_j^m(t), \tag{2.10}$$

where d_j^m is called wavelet coefficients. The decomposition is actually the correlation measure between the wavelet and the signal $X(t)$. Different frequencies are detected by varying the scale j of wavelets followed by the correlation measurement between the wavelet and data. Time localization of frequency is achieved by using index m which determines the location of frequency along the time axis.

Conversely, the function $X(t)$ can be synthesized from wavelet coefficients d_j^m by calculating

$$X(t) = \sum_{j=1}^{K} \sum_{m=0}^{2^{K-j}-1} d_j^m \phi_j^m(t), \quad 0 \le t < 2^K. \tag{2.11}$$

Armed with the two basic wavelet tools, wavelet modelling of network traffic can now be described. The wavelet traffic model was implemented as a synthetic traffic generator in [32] with the following steps

1. Parameter estimation
   - Decompose the data point into wavelet coefficients, d_j^m, using wavelet transform Eq. (2.10)
   - Calculate the empirical distribution of d_j^m at each scale (or frequency) j
2. Synthesizing

- Generate new wavelet coefficients for all scale (frequency) j with the calculated empirical distribution.
- Apply inverse transform Eq. (2.11) on generated wavelet coefficients to obtain the synthetic trace.

Examples of video traffic modelling using wavelet are discussed in [33] [34] [35].

FARIMA

The FARIMA(p,d,q), or Fractional Autoregressive Integrated Moving Average [30] [21] is an extension to the traditional ARMA(p,q) to model both SRD and LRD characteristics of traffic. Rearranging Eq. (2.8)

$$\begin{aligned}(1 - \alpha_r B - \ldots - \alpha_p B^P)X_k &= (1 - \beta_r B - \ldots - \beta_q B^q)\varepsilon_k \\ \phi(B)X_k &= \theta(B)\varepsilon_k,\end{aligned} \tag{2.12}$$

where B is the backshift operator, i.e. $BX_k = X_{k-1}$. If X_k is fractionally differenced, i.e. $\nabla^d X_k$, then

$$\begin{aligned}\phi(B)\nabla^d X_k &= \theta(B)\varepsilon_k \\ X_k &= \phi^{-1}(B)\theta(B)\nabla^{-d}\varepsilon_k,\end{aligned} \tag{2.13}$$

where $d \in (-0.5, 0.5)$ and $\nabla^d = (1 - B)^d$. FARIMA exhibits the LRD property if $d \in (0, 0.5)$. Eq. (2.13) is the FARIMA technique which has been widely used in the literature to model long-range dependent traffic. Synthesising the series X_k can be seen as filtering the fractionally differenced noise, $\nabla^{-d}\varepsilon_k$, with ARMA(p,q) [30]. Although the FARIMA model is both SRD and LRD, it suffers from computational complexity with the increasing number of data points required [23]. FARIMA was used in [9] [36] to model the video traffic characteristics.

M/G/∞

M/G/∞ [37] [38] constitutes a versatile class of models that is capable of displaying various forms of correlations by choosing an appropriate G distribution. Krunz [39] has demonstrated the capability of the M/G/∞ process for the video traffic modelling. The M/G/∞ process can be defined as follows: consider a discrete-time M/G/∞ queue in which customers arrive in IID Poisson batches of mean λ. Let ε_{n+1} be the size of the $(n+1)^{th}$ batch arrived during time slot [n,n+1). Upon arrival, the batch of customers is presented to an infinite group of servers. Customer arrivals at time [n,n+1) are serviced at the beginning of time slot [n+1,n+2). The service time of each customer in the batch is represented by integer-valued $\sigma_{n+1,1}, ..., \sigma_{n+1,\varepsilon_{n+1}}$ generated from a common distribution G. Let b_n be the number of busy servers (i.e. remaining customers) in the system at time $n = 0, 1, ...$ after counting arrivals and departures at the start of slot [n,n+1). The process $\{b_n : n = 0, 1, ...\}$ is known as the M/G/∞ input process. $\{b_n\}$ was shown to display a variety of correlation behaviour by controlling distribution G. The CDF of G has been proved to relate to the autocorrelation, $\rho(k)$, of the traffic by [39]

$$P[\sigma < k] = 1 - \frac{\rho(k) - \rho(k+1)}{1 - \rho(1)}. \tag{2.14}$$

Since $P[\sigma < k] < P[\sigma < k+1]$, it is required that $\rho(k)$ be monotonically decreasing.

2.2.4 Conclusion

In general, LRD based models are in favored over Markov or Regression based models, due to their ability to capture the autocorrelation behavior of network traffic over a long range. However, the choice of tools are dependent on the problems at hand. Also, certain tools require more computational power or more

model parameters than others. The researcher should choose the mathematical tools depending on his need and at times, combination of tools may yield a better results.

2.3 Wireless Multimedia Resource Allocation and Scheduling

This section gives an overview on topics related to wireless multimedia resource allocation and scheduling. The first part of this section describes the topics in Radio Resource Management (RRM). Secondly, the issues of multimedia resource allocation and scheduling in wireless network are discussed. The final part deals with the concept of cross layer optimisation.

2.3.1 Radio Resource Management

In this section, topics on radio resource management in wireless systems are discussed. The topic of radio resource management includes [40]

- Power Control
- Handover
- Admission and Load Control
- Link Adaptation
- Radio Resource Scheduling and Allocation

Power control

Power control refers to the process of adapting transmission power to the changing channel conditions. The varying channel conditions are due to mobility of the mobile station and interference from other base stations. Power control ensures that the radio link achieves sufficient reliability for transmission and avoids causing excessive interference to neighbouring cells. In addition, uplink power control prolongs the battery life of a small portable mobile device by transmitting at power not more than a level sufficient to maintain good link quality.

Handover

In a mobile cellular network, handover is the transition of signal transmission for a given user from one base station/sector to an adjacent base station/sector as the user moves around. To avoid denial of connection at the destination base station/sector, some system resources can be reserved to handle handover operation. Alternatively, extra resources can be obtained by degrading the QoS of existing connections to accommodate handovered connection. The resource allocation algorithm in the system should prioritise handover connection to ensure seamless change over with minimum dropped connection.

Admission and Load Control

To guarantee the QoS of ongoing connection, admission control must be adopted to determine whether to admit or reject a new connection upon its arrival. The objective of admission control is to maximize system resource utilization by admitting as many incoming connections as possible while maintaining the QoS of ongoing connections. On the other hand, load control (or congestion control) aims to manage a situation where system load exceeds the targeted threshold. Load

control may apply some counter measures, e.g. degrading the QoS of existing connections, to reduce the traffic load in order to stabilize the system.

Link Adaptation

Link adaption refers to the techniques that adapt the transmission parameters to the time varying channel by utilizing the Channel State Information (CSI). Link adaptation improves the system performances when compared to conventional non-adaptive transmission schemes.

Radio Resource Allocation and Scheduling

Radio resource allocation and scheduling are important RRM functions in mobile wireless networks. Radio resource allocation and scheduling aim to decide the amount of resource to be occupied by a mobile station and when the access to the resource should be given. The resource can be in the form of power, time slots, number of CDMA codes, frequency bandwidth, antenna etc. [41] [42] depending on the nature of air interface employed. Since resource allocation and scheduling affect the spectral efficiency of the system, vast amounts of research are required to find optimal or sub-optimal algorithms that maximize the channel utilization.

2.3.2 Issues in Wireless Resource Allocation and Scheduling

Resource allocation and scheduling problems in mobile networks are complex and multi-dimensional due to the high degree of freedom in the system parameters. Fig. 2.6 shows some of the system parameters for mobile networks. Individual parameters are discussed below.

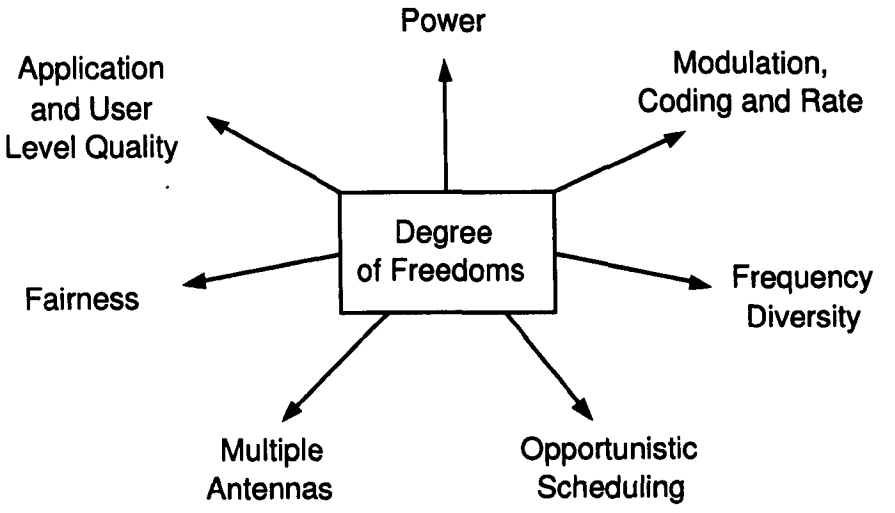


Figure 2.6: Degree of freedoms in resource allocation and scheduling

Power - Power is an important system parameter as it affects the channel quality and service coverage in the cell area. Ideally, power can be increased indefinitely to combat channel fading and noise, but this introduces interference to neighbouring cells. Also, high transmission power in the uplink direction reduces the battery life of the mobile station.

Modulation, coding and rate - Modulation, coding and rate parameters represent the modulation level, i.e. 4QAM or 16QAM, channel coding rate and sustainable data rate respectively. Joint adaptation of these parameters is referred to as Adaptive Modulation and Coding (AMC) [43]. AMC is a form of link adaptation that is used to adapt data rate to time varying channel conditions while maintaining a certain level of error performance. For example, when the channel quality is favourable, AMC increases the modulation level or reduces the channel coding redundancy to increase the link throughput and vice versa. This renders the channel bit rate time varying and complicates the resource allocation and scheduling algorithms.

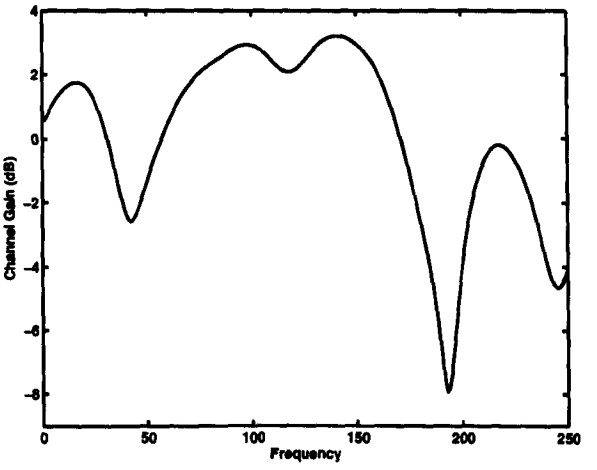


Figure 2.7: Channel gain for different frequencies

Frequency Diversity - For a system with multiple frequency bands, such as the Orthogonal Frequency Division Multiple Access (OFDMA), frequency diversity can be exploited to increase system performances [44] [45] [46]. This idea is based on the fact that wideband channel is frequency selective as shown in Fig. 2.7. Proper channel assignment can easily avoid the fading dip. It has been shown that coupling scheduling algorithms with dynamic frequency assignment enhances the system throughput.

Opportunistic Scheduling - Opportunistic scheduling or multiuser diversity [47] [48] is a novel idea proposed by David Tse to increase the system throughput. Opportunistic scheduling utilizes the fact that channel fades are independent for different mobile stations as shown in Fig. 2.8. By selecting mobile stations experiencing good channel conditions for transmission, the overall system throughput can be significantly increased. This concept forms the basis of a scheduling technique called Proportional Fair scheduling (PF) for the CDMA High Data Rate (HDR) system. However, the gain of opportunistic scheduling is limited when the

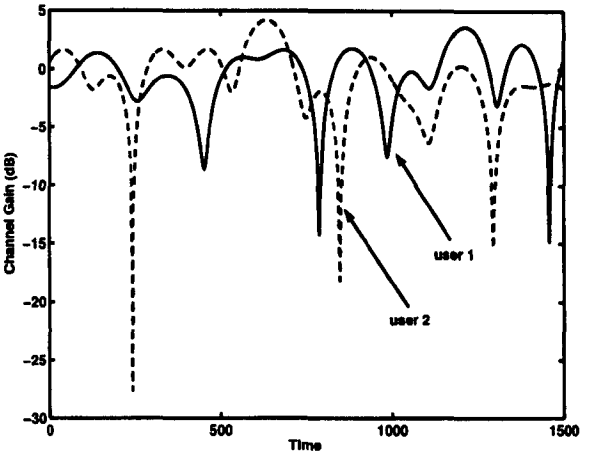


Figure 2.8: Channel gain of two users

fluctuation rate and dynamic range of channel fluctuations are small. Pramod [47] solved this problem by using an opportunistic beamforming technique to induce faster and larger channel fluctuations.

Multiple Antennas - The multiple antenna capability of modern mobile systems has led to a technique called Spatial Multiplexing (SM) [41] [49] [50] [51]. SM uses multiple transmit and multiple receive antennas for data delivery. Under rich scattering environment and sufficient antenna separation, data streams delivered on different antennas (but using the same frequency or CDMA code at the same time) can be separated at the receiver using techniques such as Zero-Forcing (ZF), Minimum Mean Square Error (MMSE) or Maximum Likelihood(ML) [52]. This greatly improves the physical layer throughput without the expense of frequency spectrum or code resource. The multiple antennas capability, or the Multiple Input and Multiple Output (MIMO) technique, is a novel idea that has revolutionized the wireless communication field. This represents another challenge or degree of freedom for resource allocation and scheduling problems as an individual antenna can be assigned to different users for transmission [53]. Another closely

related concept is Space Division Multiple Access (SDMA) [54] [55] [56]. In the SDMA system, transmit antennas are used to form a radio signal beam towards the mobile station. Since the radio signals are directed towards a certain location, interference level is low at another spatial locations. The system may reuse the same channel frequency or CDMA code for transmission to users at spatially separated locations. The simplest example of SDMA is the use of sectorization in a cell. Nevertheless, interference may be relatively high when two users occupying the same frequency or CDMA code are geographically close. It relies on the designed resource allocation or scheduling algorithms to avoid this situation e.g. allow transmission from two users on the same frequency at different time instants or on different frequencies at the same time instant.

Fairness - Fairness [57] [58] [59] [60] relates to the guaranteed throughput for a user within a time window. For example, a short-term faired system would be able to guarantee certain throughput within the time window specified by the application. Throughput guarantee in turn ensures that the packet delay is sufficiently low. As for long-term fairness, the throughput guarantee is on a longer time window, e.g. asymptotic time window, and is suitable for delay insensitve application. There are tradeoffs between short term fairness and opportunistic scheduling. An application requiring short term fairness may require access to a channel regularly. However, the scheduled user may be experiencing hostile channel conditions and transmission can only be carried out with low transmission rate. From an opportunistic scheduling point of view, this represents a loss of throughput. Algorithms should be designed to tradeoff between fairness and system throughput.

Application and User Level Quality - Another important aspect that is commonly omitted during the design of resource allocation and scheduling algorithms

is application and user level quality. For example, Wong [44] proposed an algorithm to maximize system throughput but fails to take the multimedia application quality, i.e. delay or signal distortion, into consideration. Similarly, [61] [62] give priority to user with good channel condition for transmission. Nevertheless, application level quality is usually measured by objective evaluation which may not correlate to user perceived quality. There is a growing trend to include user level quality in resource allocation/scheduling or bandwidth adaptation algorithms to enhance user perceived quality [63] [64] [65] [66].

2.3.3 Resource Allocation and Scheduling Design

Given multiple system parameters, the resource allocation and scheduling algorithms can be designed to achieve different goals as listed below

1. **Maximize system throughput** [44] [47] - A common aim of resource allocation and scheduling is to maximize the sum rate of all users given the power constraint. This is achieved by giving priority to users experiencing good channel conditions for transmission, such that higher data rate can be obtained. This method usually ignores the QoS requirements of application.

2. **Minimize transmission power** [58] - Another goal of resource allocation and scheduling algorithms is to minimize the total transmission power of a base station given the constraint of achieving certain Bit Error Rate (BER) performances. This technique utilizes the concept of opportunistic scheduling where users with good channel conditions are prioritized for transmission. As the scheduled users have good channel conditions, the sum of powers from every user is considerably less. Similarly, dynamic frequency assignment in OFDMA reduces power requirements by avoiding frequency selective fading during the assignment of frequency bands to users.

3. **Maximize sum utility** [67] - The resource allocation and scheduling algorithms in this category are application or user oriented. Resource allocation and scheduling algorithms allot system resources based on application level and user level quality. The quality is usually quantified with a utility value that is calculated by using a mathematical function. The mathematical function is pre-estimated offline by subjective evaluation to map resource allocation/scheduling decision into a satisfaction level. Thus the ultimate goal of algorithms in this category is to perform resource allocation or scheduling decisions such that the sum utility of all users is maximum.

4. **Fair scheduling** [57] - The aim of fair scheduling is to ensure bandwidth guarantee for all the users in the system considering the delay requirement. Conventionally this class of technique omits the channel conditions, but new generations of fair scheduling algorithms combine schemes such as opportunistic scheduling or beamforming to achieve better throughput performance or less transmission power.

The resource allocation and scheduling goals discussed above are common optimization targets. Due to a diverse range of system parameters and constraints, resource allocation and scheduling algorithms are complex and analytically intractable. Thus, more research is required to find reduced complexity or reduced computation algorithms for resource allocation and scheduling schemes.

2.3.4 Cross Layer Optimization

Conventional approaches adapt the transmission parameters independently, but there is a growing momentum to jointly optimize these parameters as higher performances can be achieved [68] [69]. Joint optimization is also known as cross layer optimization. The intention of this section is to give readers an overview of

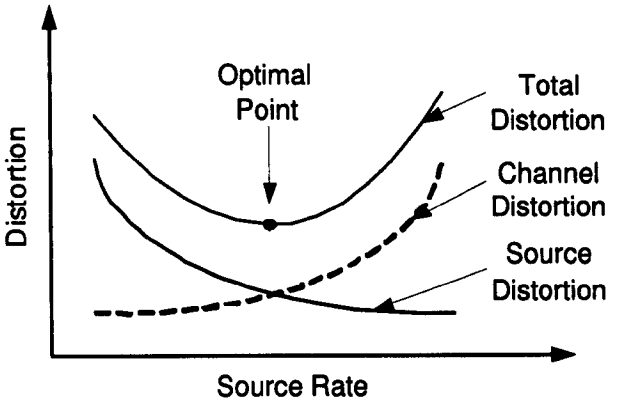


Figure 2.9: Rate-Distortion characteristic for video communication

the research topics in the area of cross layer optimization (CLO). There are three basic categories: 1) network CLO, 2) application CLO, 3) protocol CLO.

Network CLO - In this thesis, network CLO refers to the optimization of scheduling algorithm by utilizing the multiuser diversity concept. In multiuser diversity, the network CLO uses that fact that multiple users have independent channel conditions. Scheduling user with best channel quality at any particular time instant improves total channel throughput. Individual user will eventually get his fair share of throughput over the long run as the channel improves. This concept extends naturally into spatially multiplexed multiple antennas system by using the spatial diversity characteristics of MIMO system. The described concept is in fact optimization of scheduling algorithms over different protocol layers and different users to achieve high channel capacity. However, system designer must exercise care during the implementation of scheduling algorithm as certain schemes, e.g. transmit diversity, reduces the gain of network CLO [51].

Application CLO - It has been shown in the past that the schemes that jointly consider both application and transmission parameters are superior in application

level quality. Using video streaming as an example, joint optimization techniques such as Joint Source and Channel Coding (JSCC) achieve higher video quality [70] [71] [72] [73] [74] [75]. An example is shown in Fig. 2.9 for the transmission of video over a wireless channel with fixed bandwidth. In video coding, the distortions are classified into quantization distortion and channel distortion [70]. Increasing the source rate would reduce the quantization distortion. However, as the channel bandwidth is fixed, increasing the source rate reduces the amount of channel coding rate. This in turn induces higher channel distortion. There exists an optimal point where the total distortion is minimal. Joint optimization of video coding parameters (application layer) and channel coding parameters (link layer) are essential to operate at this optimal point. Another example of Application CLO is MAC layer scheduler and Automatic Repeat reQuest (ARQ). MAC layer scheduler can be made application aware i.e. aware of the semantics of the packet in the queue and prioritize transmission accordingly. ARQ operation also needs to take the packet delay requirement into consideration during re-transmission [76].

Protocol CLO - Conventional TCP assumes reliable transmission medium. Thus it is assumed that packet loss is due to congestion in the networks, and he transmission rate of the sender should be reduced. However, this assumption is not valid for wireless networks due to unreliable transmission channels [69]. In order to alleviate this problem, cross layer signalling is required so that TCP can differentiate between congestion loss or wireless channel loss. Also, retransmissions of loss data are performed by both TCP and MAC layer ARQ [69]. Some form of coordination between TCP and MAC ARQ transmission schemes are required to avoid transmitting a redundant copy of the packet.

One of the most important aspects of CLO is the cross layer signalling/dialogue.

Current internet employs layering architecture with minimal information exchange between layers. In order to enable CLO, some form of signalling/dialogue between layers is required. Selection of parameters for exchange and frequency of exchange would be vital to CLO. More discussions are available in [68] [69]. The IST PHOENIX [77] project also attempts to address these issues by proposing JSCC schemes and cross layer signalling mechanisms.

Chapter 3

Video Traffic Model With Cross Correlation Modelling

3.1 Introduction

Future communication systems are designed to support a diverse range of services. Multimedia services, especially video streaming, are foreseen to be one of the major traffic types in future communication systems. Therefore the design and evaluation of the communication systems for efficient video transmission require the characterization and modelling of video traffic. The video traffic model can be utilized as an analytical tool for queuing performance studies or as a synthetic traffic generator in software simulations.

Video traffic modelling typically involves two steps. First, the marginal distribution of the empirical frame size needs to be modelled. This captures how likely a particular frame size is present in the empirical sequence. Secondly, the AutoCorrelation Function (ACF) of the frame size sequence needs to be modelled. The ACF measures the temporal dependence between the current frame size

and previous frame sizes. More explicitly, large/small frame size tends to follow large/small frame size if the ACF value is high. The modelling of frame size ACFs can be seen as a method to capture the traffic burstiness, which has great impact on queuing performance [78]. Hence the ACF must be modelled accurately to gain precise representation of the empirical traffic characteristics. Three main classes of autocorrelation structure modelling techniques can be found in the literature: Markovian-based [26] [79], Regressive-based [28] [31] [29] [80] [81], and LRD-based [9] [36] [33] [34] [25] [17].

Rose [26] has modelled MPEG video traffic at GOP level using a simple Markov model. The GOP sequence was partitioned into scenes, and the author then classified the scenes into several scene states based on their mean GOP size in the scene. Each scene state corresponds to a Markov state. Rose has then calculated the transition probabilities between scene states. Within a single scene state, the GOP size has been divided into several GOP states and transition probabilities between GOP states have been calculated. Thus the model is composed of two nested Markov processes, where the outer scene state transitions correspond to scene changes while GOP state transitions within a scene state correspond to GOP size process about a mean GOP size. The generated GOP sizes have been converted to I, P, and B frame sizes using a method described in [26]. Sarkar *et al.* [79] have modelled the video traffic using Markov Renewal Process. They have partitioned GOP sequences into video clips and grouped the clips into seven shot classes using geometrically separated class-size boundaries. For each shot class, they modelled I, P, and B frame sizes with separate Gamma distribution. Sakar *et al.* then calculated the transition probabilities between shot classes. The sojourn time within a shot class has been modelled using Gamma distribution.

Krunz and Tripathi [28] have observed that Intra-coded (I) frame sizes fluctuate

about a mean level within a scene. Thus an Independent Identically Distributed (IID) process has been used to model the I frame size mean level. Variations of I frame size around the mean level have been modelled with an autoregressive process. Scene length duration has been fitted to Geometric distribution. For P and B frames, two separate IID processes have been used. Liu *et al.* [31] have improved Krunz and Trapathi's model by replacing the IID I frame size mean level process with an autoregressive process to account for the long range dependency (LRD). The improved model is called the Nested AutoRegressive (Nested-AR) video traffic model. Golaup and Aghvami [29] have modelled the MPEG video traffic at GOP level using an autoregressive process. H. Zhu *et al.* [80] have proposed a TES method to model the autocorrelation of I, P, and B frame sizes. Alheraish *et al.* [81] have considered the modelling of I, P, and B frame sizes using a Gaussian Autoregressive and Chi-square process (GACS). GACS is not generalized and is limited to gamma distributed frame sizes only.

Garrett and Willinger [9] have modelled the JPEG-like video traffic using a FARIMA process. As for MPEG video traffic, Ansari *et al.* [36] have modelled individual I, P, and B sequences using a separate FARIMA process. In [34] [33], Ma and Ji have modelled the MPEG video traffic at GOP level using wavelets. In [25], Jelenkovic has derived the Spatial Renewal process (SRP) for the modelling of MPEG video traffic at GOP level. The strength of SRP lies in its' capability to model any form of autocorrelation structure given that the empirical autocorrelation structure is convex and non-increasing. Huang *et al.* [17] have concentrated on modelling the autocorrelation of video VBR traffic at GOP level using a multifractal multiplicative method. The generated GOP size has been further mapped to I, P, and B frame sizes using a linear regression method. Due to the nature of the linear regression method, a frame size with negative value may be generated.

Most of the aforementioned models have ignored the cross correlation between different frame types. This underestimates the network queuing performance. In order to improve the accuracy of conventional models, the proposed video traffic models capture the cross correlation in between I, P and B frames using a Multi-noMial (MM) method. The usefulness of MM in modelling the video traffic is demonstrated in two approaches. The first approach uses a combination of MM and Spatial Renewal Process (SRP) and is called the SRP-MM model. In the second approach, the MM is utilized to enhance the existing model, the Nested-AR model. The enhanced model is called the Nested-AR-MM. For the proposed models, SRP and Nested-AR are the autocorrelation modelling technique.

The rest of this chapter is organized as follows. Section 3.2 presents the analysis of the encoded video traffic. The MM, which is used to generate a set of correlated variables, is described in Section 3.3. The procedures for modelling the marginal distribution of frame sizes are outlined in Section 3.4. Section 3.5 presents the proposed SRP-MM video traffic model, which uses MM and SRP. In Section 3.6, the MM is utilized to enhance the Nested-AR model. Model validation and performance comparison are given in Section 3.7. The conclusions are given in Section 3.8.

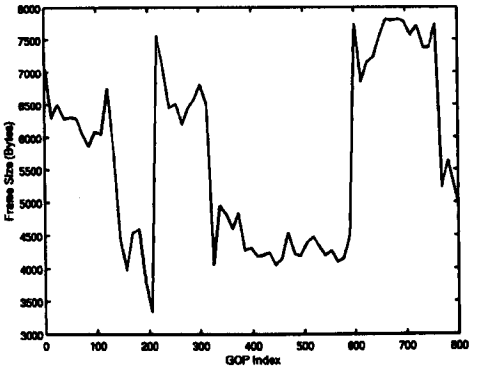
3.2 Analysis of MPEG VBR Video Traffic

In this section, the characteristics of MPEG VBR video traffic are analyzed. The MPEG4 codec is chosen for study since it plays a dominant role in various video streaming and conferencing applications. Furthermore, MPEG4 has been selected as one of the 3GPP standard codecs [8]. Nevertheless, the analysis results would apply to other codecs with similar video coding techniques, i.e. MPEG1, MPEG2

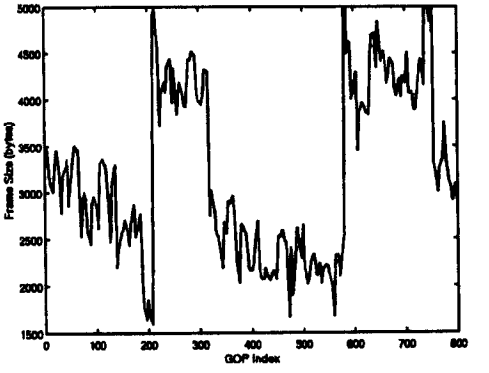
and H.264. The MPEG4 reference software from [82] is used to generate video traffic traces for analysis. There are 3 basic frame types in MPEG video encoding: I, P and B frames. An I frame is an intra-coded frame without any reference to other frames. As for a P frame, it is obtained by compressing the differential information between original frame and predicted frame where the predicted frame is estimated from a previous I or P frame. A B frame is compressed similarly to a P frame, but the predicted frame can be estimated bi-directionally from both the previous and future I or P frames. In the encoded sequence, I, P and B frames are arranged in a fixed deterministic pattern called the Group of Pictures (GOP). For example, a specific GOP arrangement can be IBPBPB or IBBPBBPBBPBB depending on the encoder parameter setting.

The MPEG4 Advanced Simple Profile (ASP) [83] is used for encoding the video sequence. A profile in MPEG context is defined as a set of tools that a decoder contains for a particular application. The MPEG4 ASP contains a set of suitable video streaming tools such as the bi-directionally predicted frames, the B frame, for enhanced compression efficiency, as well as an error resilience and error concealment tool. The readers are referred to [83] for further details of ASP. Several video sequences are used for this study, but only results from the sequences "Lord of the Rings: The Two Towers" and "Gladiator" are presented. The video sequences "Lord of the Rings: The Two Towers" and "Gladiator" have a duration of 204 minutes and 148 minutes and are respectively encoded at 25 fps and QCIF display size. The quantization parameter for each frame type is set to 4. The typical GOP pattern for video streaming, IBBPBBPBBPBB, is selected. The generated empirical frame size sequence is decomposed into separate I, P and B frame sequences for analysis.

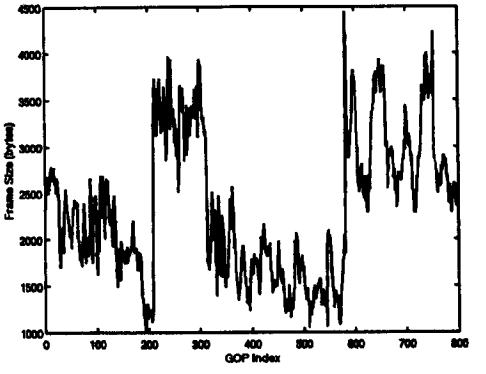
The I, P and B frame size sequences from "Lord of the Rings: The Two Towers"



(a) I frame



(b) P frame

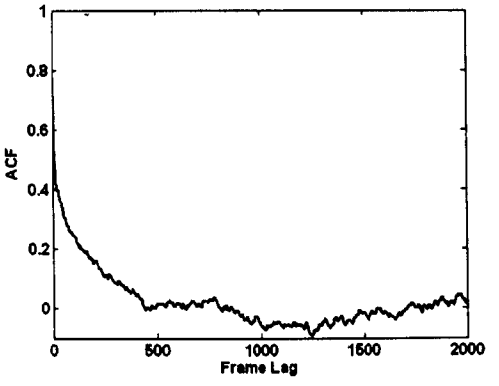


(c) B frame

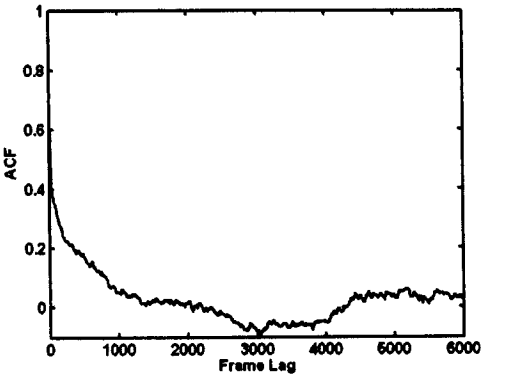
Figure 3.1: I, P and B frame size sequences

are shown in Fig. 3.1. The GOP index value is described as follows: a GOP index value of 1 refers to the first GOP, and a GOP index value of 2 refers to the second GOP. This is similar for other GOP index values. As can be observed, the mean video bit rate fluctuates noticeably from one mean level to another. This is due to scene changes in the video sequence. The duration where the mean video bit rate stays at a level is called the scene duration. This can be calculated using an algorithm described in [31]. The mean bit rate can then be obtained by calculating the average of video bit rate within the scene duration window. It can be noticed in Fig. 3.1 that the scene duration for both P and B frame sizes are predicted to be multiples of 3 and 8 to that of the I frame scene duration. As an example, one of the calculated scenes starts at GOP index 300 and ends at GOP index 600. The I frame scene duration is about 300 frames. The scene duration for P and B frames are 3x(600-300)=900 and 8x(600-300)=2400 frames. This is due to the GOP pattern IBBPBBPBBPBB where there are 3 P frames and 8 B frames in each GOP.

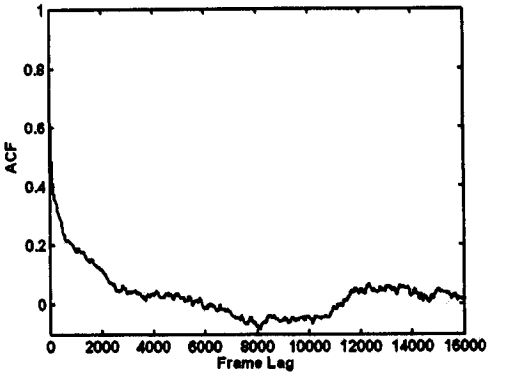
Fig. 3.1 shows the dependency between the I, P and B frame sizes where I, P and B frame sizes follow the same trend. Furthermore, it can be seen in Fig. 3.2 that the ACFs of both P and B frame sizes have a similar shape to the I frame sizes ACF, but only with dilated lag index at multiples of 3 and 8 of the I lag index. This is expected since I, P and B frame sizes follow a similar trend (cross correlated) and thus have a similar ACF shape. The scene duration of P and B frames is at multiples of I frame scene duration which causes a dilated lag for P and B frame sizes ACFs. All of these observations suggest that the ACFs of P and B frame sizes can be predicted from the ACF of I frame sizes.



(a) I frame



(b) P frame



(c) B frame

Figure 3.2: I, P and B frame size AutoCorrelation Function (ACF)

3.3 MultinoMial Method (MM)

The proposed traffic model considers the cross correlation between different frame types to increase the accuracy of the conventional model. The cross correlation between different frame types (I, P, and B frame sizes) is modelled using a MultinoMial method (MM). Table 3.1 presents the cross correlation matrix for I, P and B frame sizes. It can be observed that the correlation coefficient can be as high as 0.9396. The correlation coefficient between I, P and B frame sizes is measured using the cross correlation function

$$\begin{aligned} XCorr(J,K) &= \frac{Cov(J,K)}{\sqrt{Cov(J,J) \times Cov(K,K)}} \\ &= \frac{E[(J-\mu_J)(K-\mu_K)]}{\sigma_{JJ} \times \sigma_{KK}}, \end{aligned} \tag{3.1}$$

where Cov is covariance and $\sigma_{JJ} = \sqrt{Cov(J,J)}$. J and K are two different time series whereas μ_J and μ_K are the mean of series J and K respectively. For the purpose of cross correlation measurement, all the data points for P and B frame types within a GOP ($I_1B_1B_2P_1B_3B_4P_2B_5B_6P_3B_7B_8$) are summed and averaged respectively to form a new GOP sequence $I_1B_{avg}P_{avg}$ so that I_1, B_{avg} and P_{avg} have an equal number of data points. This assumes that P and B frame sizes within a GOP have little variation in value and the averaging process has an insignificant effect on the model performance. The MultinoMial method [84] is

Frame Type	I	P	B
I	1.0000	0.7058	0.6190
P	0.7058	1.0000	0.9396
B	0.6190	0.9396	1.0000

Table 3.1: Correlation matrix between frame types

described as follows: Let the covariance matrix between I, P and B frames be

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 \end{pmatrix}, \tag{3.2}$$

where σ_{ij}^2 is covariance between i and j variables. Without loss of generality, assume that $\sigma_{ii}^2 = \sigma_{jj}^2 = 1$. Hence from Eq. (3.1),

$$\begin{aligned} \sigma_{ij}^2 &= Cov(i,j) \\ &= XCorr(i,j) \times \sigma_{ii} \times \sigma_{jj} \\ &= XCorr(i,j). \end{aligned} \tag{3.3}$$

Hence the coefficients of the covariance matrix in Eq. (3.2) can readily be obtained from Table 3.1 using Eq. (3.3). Suppose that a vector $\mathbf{X} = (X_1, X_2, X_3)$ be generated as using a linear combination of Gaussian distributed variables $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$

$$\mathbf{X} = \mathbf{LZ}, \tag{3.4}$$

where $\mathbf{L}$ is a 3×3 matrix and all elements of vector $\mathbf{Z}$ are random Gaussian variables with zero mean and unity variance. It has been proved that $\mathbf{X}$ is Gaussian distributed with zero mean and covariance matrix $\mathbf{LL^T}$ (i.e. $N(0, \mathbf{LL^T})$) [84]. It is desired that the covariance of $\mathbf{X}$ which is $\mathbf{LL^T}$ be equal to empirically measured covariance $\mathbf{\Sigma}$. Thus,

$$\mathbf{\Sigma} = \mathbf{LL^T}. \tag{3.5}$$

Assuming that $\mathbf{\Sigma}$ from Eq. (3.2) is symmetric and positive definite, then $\mathbf{\Sigma}$ can be factorized into $\mathbf{L}$ using the Cholesky Decomposition [85]. The factorized $\mathbf{L}$ has the form of

$$\mathbf{L} = \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix}. \tag{3.6}$$

Eq. (3.5) can be rewritten as

$$\begin{aligned}\begin{pmatrix}\sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2\\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2\\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2\end{pmatrix} &= \begin{pmatrix}L_{11} & 0 & 0\\ L_{21} & L_{22} & 0\\ L_{31} & L_{32} & L_{33}\end{pmatrix}\begin{pmatrix}L_{11} & L_{21} & L_{31}\\ 0 & L_{22} & L_{32}\\ 0 & 0 & L_{33}\end{pmatrix}\\ &= \begin{pmatrix}L_{11}^2 & L_{11}L_{21} & L_{11}L_{31}\\ L_{11}L_{21} & L_{21}^2+L_{22}^2 & L_{21}L_{31}+L_{22}L_{32}\\ L_{11}L_{31} & L_{21}L_{31}+L_{22}L_{32} & L_{31}^2+L_{32}^2+L_{33}^2\end{pmatrix}.\end{aligned} \tag{3.7}$$

Using expression from Eq. (3.3), the equation above can be simplified to

$$\begin{pmatrix}1 & \rho_{12} & \rho_{13}\\ \rho_{12} & 1 & \rho_{23}\\ \rho_{13} & \rho_{23} & 1\end{pmatrix} = \begin{pmatrix}L_{11}^2 & L_{11}L_{21} & L_{11}L_{31}\\ L_{11}L_{21} & L_{21}^2+L_{22}^2 & L_{21}L_{31}+L_{22}L_{32}\\ L_{11}L_{31} & L_{21}L_{31}+L_{22}L_{32} & L_{31}^2+L_{32}^2+L_{33}^2\end{pmatrix}, \tag{3.8}$$

where $\rho_{ij} = XCorr(i,j)$. Solving the equation above yields

$$\begin{aligned} L_{11} &= 1 & L_{22} &= \sqrt{1-L_{12}^2} & L_{32} &= \sqrt{\frac{\rho_{23}-L_{21}L_{31}}{L_{22}}}\\ L_{21} &= \rho_{12} & L_{31} &= \rho_{13} & L_{33} &= \sqrt{1-L_{32}^2-L_{31}^2}. \end{aligned}$$

Expanding Eq. (3.4) using Eq. (3.6) and performing variance normalization yields

$$\begin{aligned} X_1^N &= \frac{L_{11}Z_1}{\sqrt{L_{11}^2}}\\ X_2^N &= \frac{L_{21}Z_1+L_{22}Z_2}{\sqrt{L_{21}^2+L_{22}^2}}\\ X_3^N &= \frac{L_{31}Z_1+L_{32}Z_2+L_{33}Z_3}{\sqrt{L_{31}^2+L_{32}^2+L_{33}^2}}. \end{aligned} \tag{3.9}$$

The individual element of vector $\mathbf{X^N} = (X_1^N, X_2^N, X_3^N)$ is a zero mean and unity variance Gaussian variable and can be respectively mapped to I, P and B frame sizes using probability integral transform Appendix B.1.

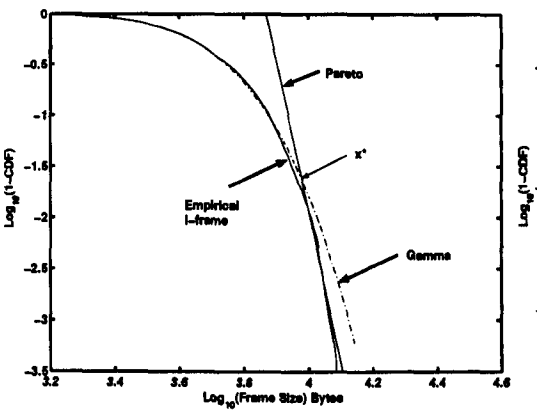
3.4 Frame Size Marginal Distribution Modelling

The frame size (expressed in total number of bytes) marginal distribution captures how likely a particular frame size is to be present in a video sequence. The frame size characteristics may change when, for example, a different quantization parameter is used, or when a different video sequence with different content complexity and motion activity is used or when a different resolution is used. Since the frame size characteristics of encoded video traffic is eventually reflected in the marginal distribution, a traffic model which is based on the marginal distribution modelling can be used to predict video traffic characteristics for any video sequences, codec type and the encoder parameters.

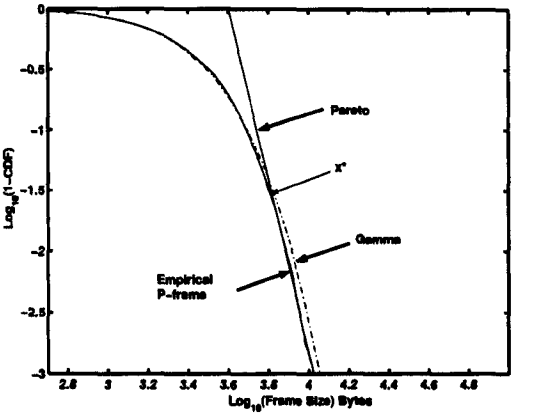
The marginal distribution of I, P, B frame sizes has been modelled using a hybrid approach proposed in [9]. The Gamma and Pareto distributions are respectively found to capture the general body shape and tail of empirical I, P and B frame sizes Cumulative Distribution Function (CDF) accurately. Fig. 3.3 shows the fitted frame size distribution for the empirical trace. Let F_Γ and F_P denote the CDF for Gamma and Pareto distributions; the hybrid Gamma/Pareto distribution that is used to fit the empirical distribution is given by

$$F_{\Gamma/P}(x) = \begin{cases} F_\Gamma(x), & \text{if } x \leq x^* \\ F_P(x), & \text{if } x > x^* \end{cases}, \tag{3.10}$$

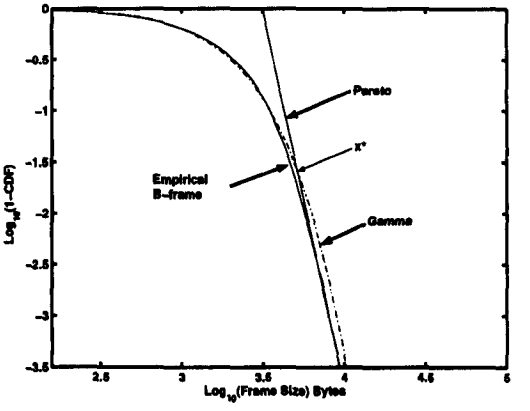
where x^* is the cut-off point when the empirical distribution starts to deviate from the Gamma fit as shown in Fig. 3.3. Once x^* is determined, the Pareto distribution can be fitted along the empirical distribution tail using least square fitting. The Gamma and Pareto distributions are described in Appendix B.1.



(a) I frame



(b) P frame



(c) B frame

Figure 3.3: I, P and B frame size marginal distribution fitting

3.5 Spatial Renewal Process and MultinoMial Video Traffic Model

This section describes the proposed Spatial Renewal Process and MultinoMial (SRP-MM) video traffic model.

3.5.1 Spatial Renewal Process (SRP)

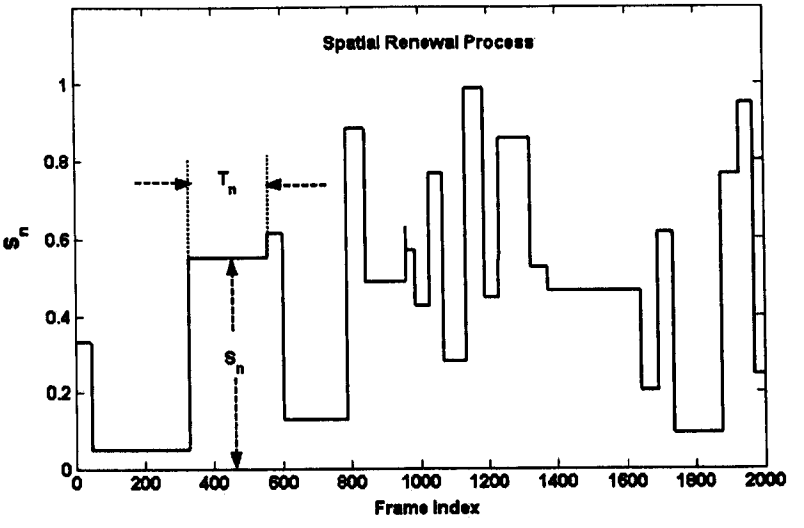


Figure 3.4: Spatial Renewal Process sample path

The SRP [25] is used to model the ACFs of frame sizes. The SRP is chosen due to its low computational complexity requirement. The SRP process, $Y(t)$, is composed of a chain of renewal periods, where the n^{th} period is T_n in length, and the sample path during this period takes on the value S_n. The inter-renewal time $\{T_n\}$ and sample path value $\{S_n\}$ are IID. An example of the SRP sample path is depicted in Fig. 3.4. In the video traffic modelling context, the level, S_n, can be regarded as the mean video bit rate during a scene, and the inter-renewal

time, T_n, can be regarded as the scene duration. The mean video bit rate level shifting can be attributed to scene changes. It has been shown in [25] that the ACF of SRP, $\rho(\tau)$, is related to the inter-renewal time distribution, $F_T(\tau)$ by

$$F_T(\tau) = 1 - \frac{\rho(\tau) - \rho(\tau+1)}{1-\rho(1)}. \tag{3.11}$$

Note that since $F_T(\tau) \leq 1$, $\rho(\tau)$ is necessarily convex and non-increasing. For all the video sequences considered, the I frame sizes ACF is used to estimate $F_T(\tau)$. Therefore, the generated T_n represents the number of I frames that stays at a constant mean bit rate level in the n^{th} scene. The total number of P and B frames in the n^{th} scene are respectively 3 and 8 times the total number of I frames, as described in Section 3.2.

The I frame sizes ACF curve is found to be well fitted to the function $e^{-b\sqrt{\tau}}$ where b is the fitted parameter from the ACF of I frame sizes. The time distribution $F_T(\tau)$ is then obtained from Eq. (3.11) as

$$F_T(\tau) = 1 - \frac{e^{-b\sqrt{\tau}} - e^{-b\sqrt{\tau+1}}}{1-e^{-b}}. \tag{3.12}$$

3.5.2 SRP-MM Video Traffic Model

The SRP and MM, as discussed in previous sections, are used to model the encoded video traffic. The complete SRP-MM model is shown in Fig. 3.5. The synthetic frame size generation process is summarized as follows:

1. Generate Z_1 as zero mean and unity variance Gaussian variable using SRP. Then generate Z_2 and Z_3 randomly as zero mean and unity variance Gaussian variable.

2. Use MM to generate the correlated vector $\mathbf{X^N} = (X_1^N, X_2^N, X_3^N)$ using Z_1, Z_2, Z_3 and Eq. (3.9).

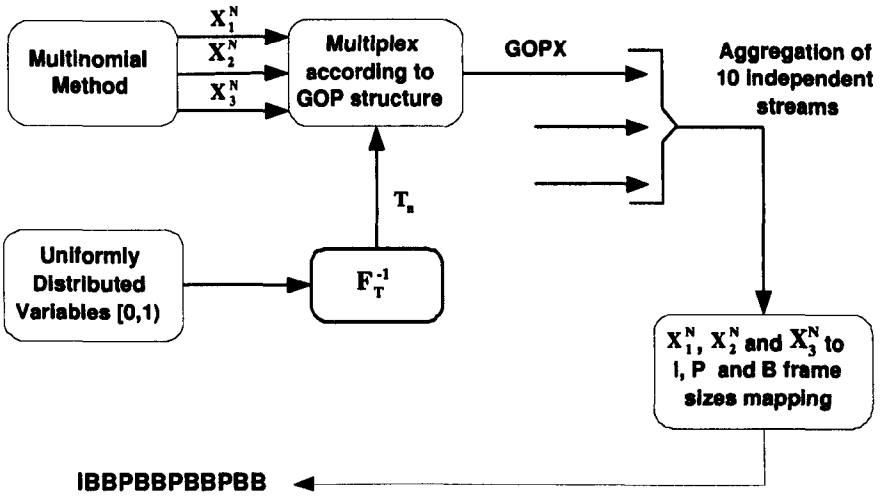


Figure 3.5: SRP-MM video traffic model

3. Generate T_n from Eq. (3.12). As explained in Section 3.5.1, there are T_n I frames, $3 \times T_n$ P frames and $8 \times T_n$ B frames in the scene. Arrange X_1^N, X_2^N, and X_3^N in an appropriate GOP structure. The GOP vector is denoted as **GOPX**.

4. Map X_1^N, X_2^N, and X_3^N to I, P and B frame sizes using the probability integral transform described in Appendix B.2. Take the mapping of X_1^N to I frame size for example. The probability integral transform theorem states that for any given distribution function $F(x)$

$$U = F(x), \tag{3.13}$$

is uniformly distributed where $F(x)$ is the cumulative probability at x. Therefore the Gaussian distribution, $F_G(X_1^N, \mu, \sigma)$, and the hybrid Gamma/Pareto distribution, $F_{\Gamma/P}(S_I)$, can be equated since they are both uniformly distributed,

$$\begin{aligned} F_{\Gamma/P}(S_I) &= F_G(X_1^N, \mu, \sigma) \\ S_I &= F_{\Gamma/P}^{-1}(F_G(X_1^N, \mu, \sigma)), \end{aligned} \tag{3.14}$$

where S_I is the calculated I frame size and $F^{-1}_{\Gamma/P}$ is the inverse of Eq. (3.10).

5. Repeat steps 1 to 4 until the total number of required frames sizes are obtained.

Note that in steps 3, 10 independently generated **GOPX** processes aggregated to improve the empirical stability and variability of output sequence. Besides, the aggregation of independent homogenous streams does not destroy the ACF property [6].

3.6 Nested AutoRegressive and MultinoMial Video Traffic Model

This section introduces the Nested AutoRegressive model and the use of MM to improve the accuracy of original Nested-AR model [31].

3.6.1 Nested AutoRegressive (Nested-AR)

In [31], the empirical frame trace is decomposed into separate I, P and B frame sequences for modelling. The Nested-AR process is then introduced to model the autocorrelation of I frame sizes. The P and B frame sizes are assumed to be IID random variables. The cross correlations between different frame types are assumed to be insignificant. I, P and B frame sizes are fitted to their marginal distribution respectively.

In the Nested-AR model, $X_I(n)$, the process which models the I frame sizes is the sum of two independent normal random variables

$$X_I(n) = M_I(n) + \delta_I(n), \tag{3.15}$$

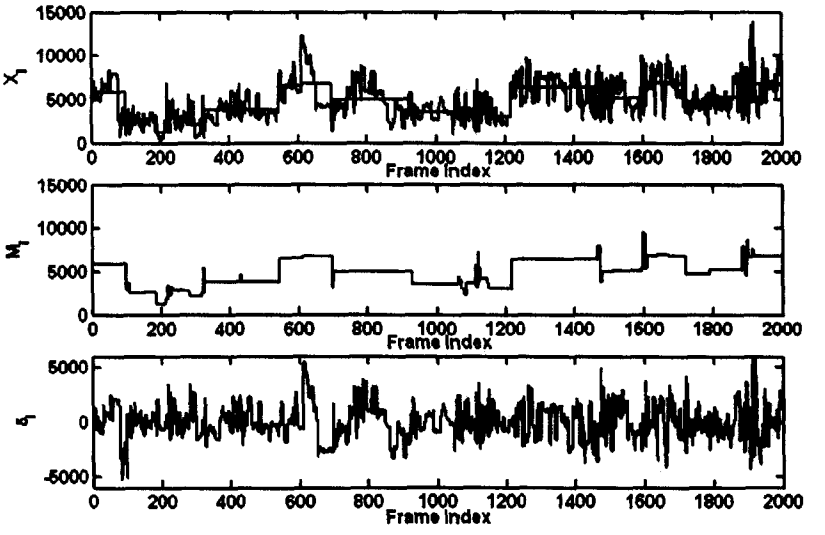


Figure 3.6: Nested AutoRegressive (Nested-AR) I frame size modelling

where $M_I(n)$ and $\delta_I(n)$ is the mean frame size and fluctuation of the frame size about its mean for the n^{th} I frame. Fig. 3.6 shows the $M_I(n)$ and $\delta_I(n)$ of the empirical trace. Scene changes are detected using the same approach described in [31]. During j^{th} scene with scene length duration N_j which starts with k^{th} I frame, $M_I(n)$ has the same value for every frame within the scene, which is denoted $\tilde{X}_I(j)$, i.e.

$$M_I(k) = M_I(k+1) = ... = M(k+N_j+1) \triangleq \tilde{X}_I(j). \tag{3.16}$$

In essence the $M_I(n)$ can be modelled by using the scene length duration, N_j, and mean value of the scene, $\tilde{X}_I(j)$. In the Nested-AR model, N_j is modelled with the geometric distribution. The $\tilde{X}_I(j)$ is modelled using a second order autoregressive process

$$\tilde{X}_I(j) = b_1\tilde{X}_I(j-1) + b_2\tilde{X}_I(j-2) + \theta_I(j), \tag{3.17}$$

where b_1 and b_2 are real constants and $\{\theta_I(j)\}$ is an IID normal distributed

random variables. The mean, μ_θ, and variance, σ_θ^2, of $\{\theta_I(j)\}$ are determined as

$$\mu_\theta = (1 - b_1 - b_2)\tilde{\mu}_I, \tag{3.18}$$

and

$$\sigma_\theta^2 = \frac{(1 + b_2)\left[(1 - b_2)^2 - b_1^2\right]\tilde{\sigma}_I^2}{1 - b_2}, \tag{3.19}$$

where $\tilde{\mu}_I$ and $\tilde{\sigma}_I^2$ are the mean and variance of $\tilde{X}_I(j)$. The second random variable $\delta_I(n)$ has been modelled with a second order autoregressive process

$$\delta_I(n) = a_1\delta_I(n-1) + a_2\delta_I(n-2) + \varepsilon_I(n), \tag{3.20}$$

where a_1 and a_2 are real constants and $\{\varepsilon_I(n)\}$ is IID normal distributed random variable. The mean of $\{\varepsilon_I(n)\}$ is defined as zero and its variance is

$$\sigma_{\varepsilon_I}^2 = \frac{(1 + a_2)\left[(1 - a_2)^2 - a_1^2\right]\sigma_{\delta_I}^2}{1 - a_2}, \tag{3.21}$$

where $\sigma_{\delta_I}^2$ is variance of $\delta_I(n)$. Note that $\{X_I(n)\}$ has a different marginal distribution from empirical I frame sizes marginal distribution. This is because $\{M_I(n)\}$ and $\{\delta_I(n)\}$ are normal distributed since $\{\theta_I(n)\}$ and $\{\varepsilon_I(n)\}$ in Eq. (3.17) and Eq. (3.20) are normal distributed. $X_I(n)$ needs to be mapped to I frame sizes using similar procedures to step 4 of SRP-MM in Section 3.5.2.

3.6.2 Nested-AR-MM Video Traffic Model

It can be observed in Fig. 3.1 that P and B frame sizes are autocorrelated. Furthermore, it is observed that there exist cross correlations between I, P and B frame sizes. These observations are in contrast to the assumptions made in the Nested-AR model. Based on these observations, two improvements are proposed:

1. **Model the autocorrelation of P and B frame sizes**, using the same method which is used for I frame sizes. However, scene detection is not performed on P and B frame sizes. The scene boundaries are obtained by properly scaling the I frame size scene edges.

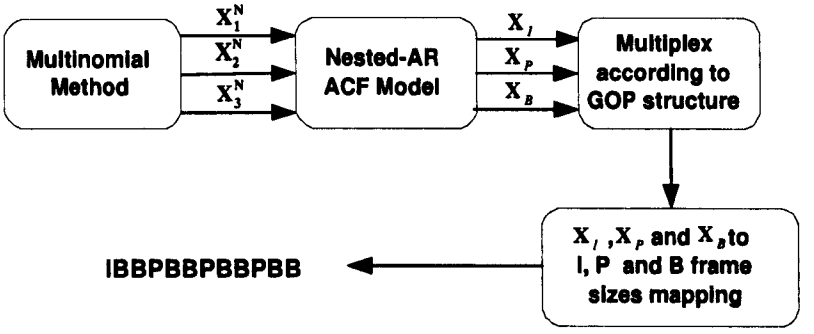


Figure 3.7: Nested-AR-MM video traffic model

2. **Model the cross correlation between different frame types.** This is achieved by modelling the cross correlation of the scene mean value of I, P and B frames.

In order to consider the P and B frame sizes ACF, the mean level of P and B frame sizes of j^{th} scene are modelled similarly to Eq. (3.17)

$$\tilde{X}_P(j) = b_1\tilde{X}_P(j-1) + b_2\tilde{X}_P(j-2) + \theta_P(j) \tag{3.22}$$

$$\tilde{X}_B(j) = b_1\tilde{X}_B(j-1) + b_2\tilde{X}_B(j-2) + \theta_B(j). \tag{3.23}$$

In view of Eq. (3.17), Eq. (3.22) and Eq. (3.23), $\tilde{X}_I(j)$, $\tilde{X}_P(j)$ and $\tilde{X}_B(j)$ can be made correlated if $\theta_I(j)$, $\theta_P(j)$ and $\theta_B(j)$ are correlated. The **MultinoMial method (MM)** can be used to generate three correlated variables X_1^N, X_2^N, and X_3^N of **zero mean** and **unity variance**. X_1^N, X_2^N, and X_3^N can be respectively mapped to $\theta_I(j)$, $\theta_P(j)$ and $\theta_B(j)$ by

$$\begin{aligned} \theta_I(j) &= \mu_{\theta_I} + \sigma_{\theta_I} X_1^N \\ \theta_P(j) &= \mu_{\theta_P} + \sigma_{\theta_P} X_2^N \\ \theta_B(j) &= \mu_{\theta_B} + \sigma_{\theta_B} X_3^N, \end{aligned} \tag{3.24}$$

where μ_{θ_I} and σ_{θ_I} are defined for I frame sizes in Eq. (3.18) and Eq. (3.19). The mean and variance of P and B frame sizes are defined similarly as μ_{θ_P}, σ_{θ_P}, μ_{θ_B} and σ_{θ_B}. The improved model is called the Nested-AR-MM, shown in Fig. 3.7.

3.7 Model Validation

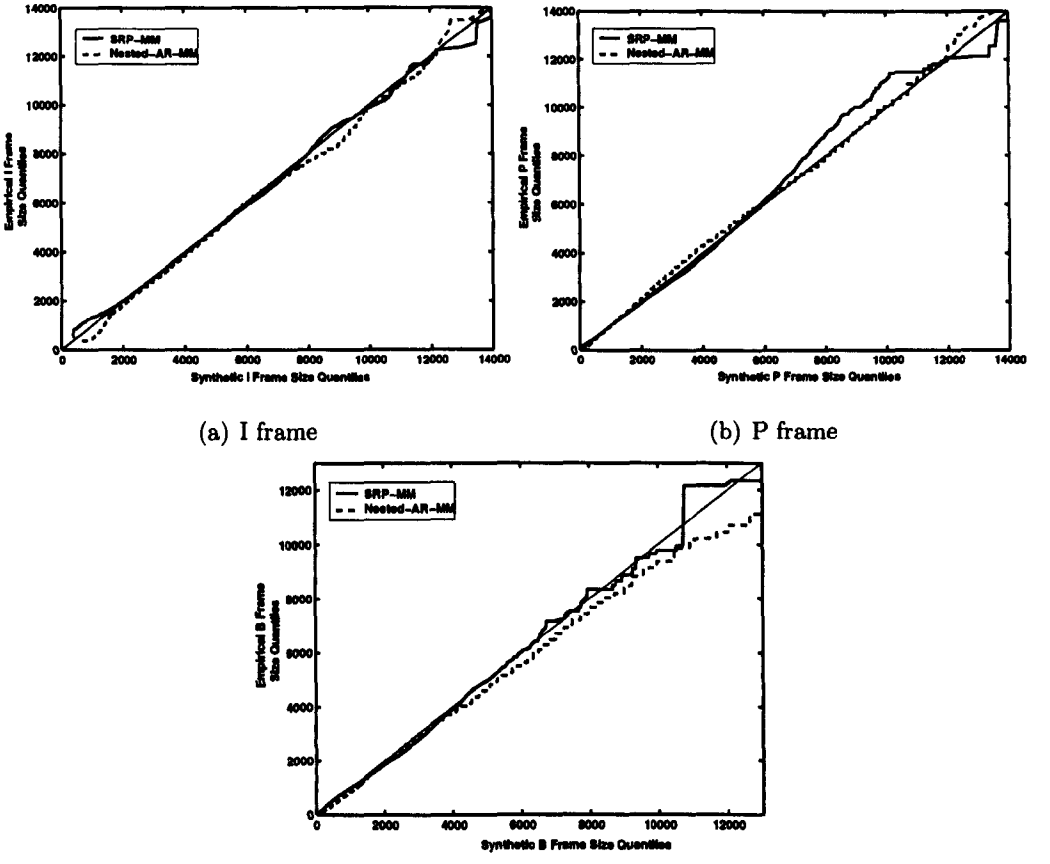
Three important techniques are considered in the validation process. They are the frame size marginal distribution, frame sizes ACF and packet loss rate prediction accuracy. In addition, the Nested-AR [31] and FARIMA [36] models are implemented for performance comparison to the proposed SRP-MM and Nested-AR-MM during the validation process.

3.7.1 Marginal distribution

The Quantile to Quantile (QQ) plot [79] is used to verify if the model can predict the marginal distribution of empirical trace accurately. The quantile of the empirical frame sizes is plotted against the quantile of the model generated frame sizes. Fig. 3.8 shows the QQ plot of the empirical frame sizes against the SRP-MM generated frame sizes for "Lord of the Rings: The Two Towers". Similar results are found for SRP-MM, Nested-AR-MM, Nested-AR and FARIMA models.

3.7.2 Frame Size ACF

The ACFs of the empirical trace and the synthetically generated I, P and B frame sizes for all the models, SRP-MM, Nested-AR-MM, Nested-AR and FARIMA are plotted in Fig. 3.9. As can be seen from the figure, the ACFs of synthetic frame sizes for all the models predict the empirical frame size ACF accurately. The exceptions are P and B frame sizes ACFs for the Nested-AR model which have zero values over the frame lag axis due to the IID assumption.

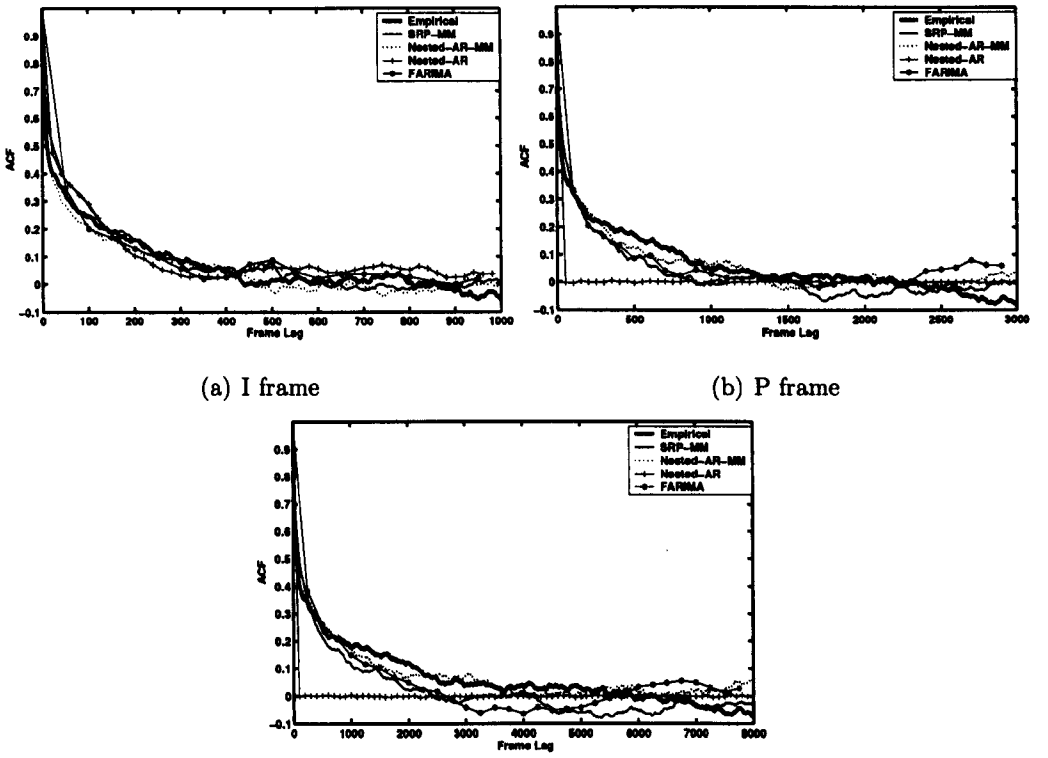


(a) I frame

(b) P frame

(c) B frame

Figure 3.8: QQ plot of empirical frame sizes against synthetic frame sizes



(a) I frame

(b) P frame

(c) B frame

Figure 3.9: I, P and B frame size ACF fitting

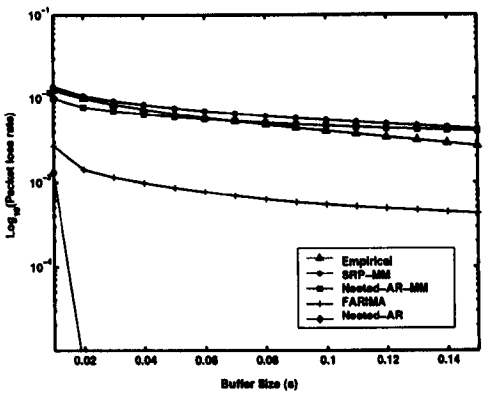
3.7.3 Packet Loss Rate Prediction

The SRP-MM and Nested-AR-MM are validated by means of packet loss rate prediction accuracy. The packet loss refers to packet dropped due to buffer overflow at a bottlenecked router. The validation by packet loss rate prediction is used to ensure that the model is capable of capturing the intrinsic burstiness of encoded VBR video traffic due to the autocorrelation and cross correlation of I, P and B frame sizes. Taking cross correlation between I, P and B frames sizes for example, large I frame sizes are likely to cluster with large P and B frame sizes in a GOP and vice versa, as shown in Fig. 3.1. If I, P and B frames are generated independently, this "clustering" is lost and the output video traffic will be less bursty. This will then cause the packet loss rate of such model to be lower when compared to their empirical counterpart.

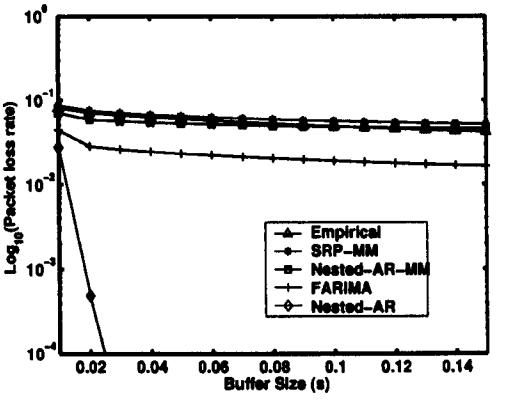
The model validation is achieved by transmitting model generated video traffic to a First In First Out (FIFO) queue. The packet loss rate of synthetic traffic is recorded and compared to the packet loss rate of empirical video traffic under the same simulation settings. The simulation is repeated for different bandwidth utilizations (U). The buffer size of the queue under study is in the range of 10 ms - 150 ms. Buffer size (D) (in seconds) is calculated by

$$D = \frac{Z}{W}, \tag{3.25}$$

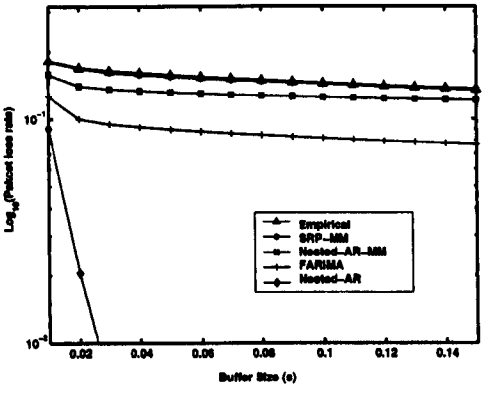
where Z is the buffer size (bits) and W is the output bandwidth (bits/second). Simulations for bandwidth utilization of U = 40%, U = 60% and U = 80% are carried out. The SRP-MM and Nested-AR-MM are also used to model several other sequences. However, only representative results based on the film sequences, "Lord of the Rings: The Two Towers" and "Gladiator" are presented in Fig. 3.10 and Fig. 3.11. It can be seen that the SRP-MM and Nested-AR-MM packet loss rate closely follow the empirical ones. It is further shown that the SRP-



(a) Bandwidth utilization 40%

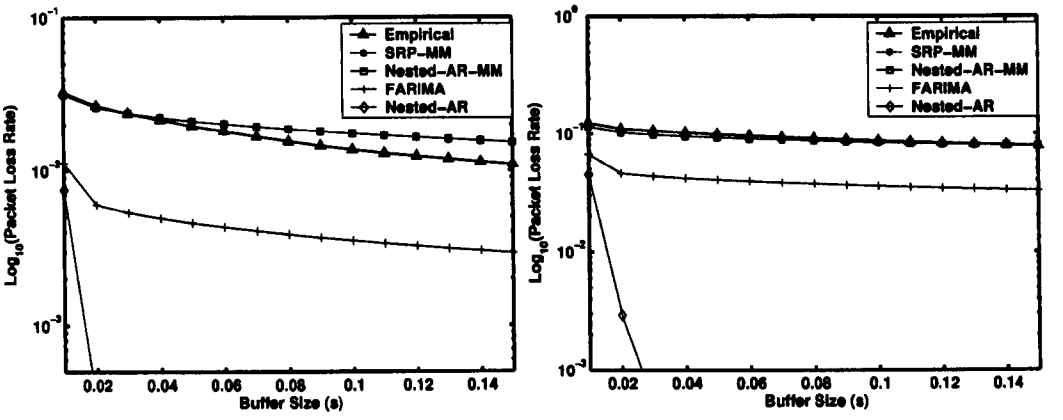


(b) Bandwidth utilization 60%



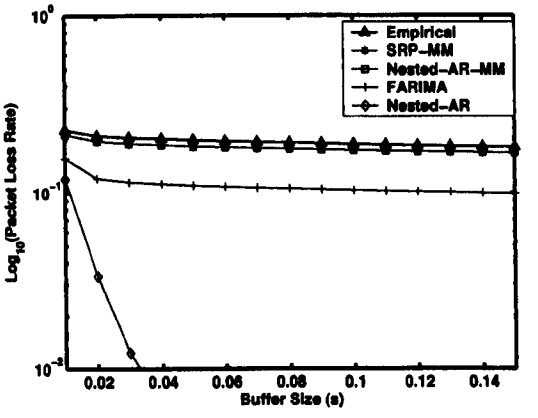
(c) Bandwidth utilization 80%

Figure 3.10: Empirical packet loss rate matching for Lord of the Rings for different bandwidth utilization



(a) Bandwidth utilization 40%

(b) Bandwidth utilization 60%



(c) Bandwidth utilization 80%

Figure 3.11: Empirical packet loss rate matching for Gladiator for different bandwidth utilization

MM and Nested-AR-MM outperform the Nested-AR and FARIMA model in terms of packet loss rate prediction accuracy. This is expected since Nested-AR and FARIMA ignored the cross correlation between different frame types. It is further found that ignoring the autocorrelation structure in P and B frames can cause a great underestimation of the packet loss rate. This can be observed in Fig. 3.10 and Fig. 3.11 where the packet loss rate of Nested-AR model deviates from the empirical curve significantly. The FARIMA model, on the other hand, has better prediction accuracy compared to the Nested-AR model since it models the autocorrelation structure of P and B frames.

3.8 Conclusions

The statistical characteristics of MPEG VBR encoded video sequence have been studied in this chapter. It is found that I, P and B frame sizes are autocorrelated and cross correlated. A MultinoMial method is proposed to capture the correlation between frame types. The usefulness of the MultinoMial method has been demonstrated in two separate approaches. The first approach uses a combination of the MultinoMial method and Spatial Renewal Process to model the MPEG VBR video traffic. The second approach uses the MultinoMial method to enhance the accuracy of the existing Nested-AR model. The proposed models, SRP-MM and Nested-AR-MM are shown to capture the marginal distribution and ACF of empirical frame sizes accurately. Simulation results show that SRP-MM and Nested-AR-MM predicts the empirical queuing performance with a high accuracy under realistic buffer sizes and different bandwidth utilizations. Results have shown that ignoring the cross correlation between frame types, as in the FARIMA and Nested-AR models, can cause an underestimation of the queuing performance due to reduced traffic burstiness when compared to the empirical trace.

Chapter 4

Generalized Video Traffic Model

4.1 Introduction

It has been demonstrated in previous chapter that SRP-MM and Nested-AR-MM can capture MPEG4 encoded VBR video with high accuracy. However, SRP-MM, Nested-AR-MM and conventional models [9] [26] [79] [28] [31] [29] [80] [81] [36] [33] [25] [17], are proposed for a fixed quantization parameter set. To put it differently, these models are designed for VBR video traffic where some fixed quantization parameter set is assumed. As the encoder parameter is varied, the model parameters need to be re-estimated from newly generated traces. Model parameter re-estimation is time consuming and the statistical properties of empirical traces may change. Under this condition, the models studied under one set of quantization parameter may no longer be valid for other sets of quantization parameter. Due to the nature of VBR traffic, these models may not be suitable for study of video streaming over wireless environments. This is because the bandwidth of wireless channels is time-varying but the bit rate of conventional models does not scale easily to match channel bandwidth as a fixed quantization parameter set is assumed. Therefore, a new video traffic model with

the capability to adapt its output traffic characteristics according to quantization parameters in real time is required. This section studies an MPEG4 video traffic model with real time adaptation capability. Adaptation capability is achieved by designing a convenient frame size model that considers the autocorrelation structure, cross correlation between different frame types and the marginal distribution of empirical frame sizes.

The rest of this chapter is organized as follows. The terminologies and fundamental concepts employed in Generalized Video Traffic Model (GVTM) are explained in Section 4.2. The modelling of frame sizes using the frame activity concept is discussed in Section 4.3. The MultinoMial (MM) method for the modelling of cross correlation between frame types is presented in Section 4.4. The Spatial Renewal Process (SRP) which is used to model the autocorrelation structure of video sequences is described in Section 4.5. A summary of GVTM traffic generation is given in Section 4.6. Performance evaluation of the proposed model and its comparison to existing models are presented in Section 4.7. The chapter is concluded in Section 4.8.

4.2 MPEG4 Encoded Video

In this section, the terminologies and fundamental concepts employed in GVTM are introduced. Section 4.2.1 introduces the frame activity concept which is used to characterize a video frame. In Section 4.2.2, the MPEG4 frame composition and its contribution to the total number of output bits are studied. The measurement method used is presented in Section 4.2.3.

4.2.1 Frame Activity

Frame activity measures the amount of detail in a frame; a higher frame activity will generate a larger output frame size (more accurately texture bits) during the encoding operation. Common methods [86] of measuring the frame activity include variance-, gradient-, DCT- and Edge-based methods. The DCT-based method is adopted for activity measurement. The DCT-based method is described as follows: the input image is first partitioned into 8×8 pixel blocks, then two dimensional DCT is performed on all the blocks. Let $a(i)$ be the average of absolute sum of DCT AC coefficients for i^{th} block, i.e.

$$a(i) = \frac{1}{63} \sum_{j=2}^{64} |DCT_i(j)|, \tag{4.1}$$

where $DCT_i(j)$ is the j^{th} DCT coefficient of the i^{th} block. The frame activity is defined as:

$$\hat{A} = \sum_{i=1}^{M} a(i), \tag{4.2}$$

where M is the total number of 8×8 blocks in a frame.

4.2.2 MPEG4 Video Frame Composition

An MPEG4 encoder generates three types of Video Object Planes (VOPs): Intra-coded VOPs (I), forward motion predicted VOPs (P) and bidirectionally motion predicted VOPs (B). For simplicity reasons, a VOP is assumed to be a normal rectangular video frame. A video frame consists of three basic components including one luminance component representing brightness and two chrominance components representing colour information. The number of bits needed after encoding for a luminance and two chrominance components are respectively represented as Y, Cr, and Cb. The total number of bits for an I frame comprises header bits (H), luminance bits (Y), chrominance bits (Cr and Cb). For simplicity, H, Y, Cr, and Cb are jointly considered as TeXture bits (TX). For P and B

frame types, the total number of bits consists of TeXture bits (TX) and Motion Vector bits (MV). MV is a result of motion estimation algorithm for inter-frame predictions [83]. TX depends on the frame activity (or amount of details in a frame), and also the quantization parameter. For example, larger frame activity yields larger texture bits, but using a large quantization parameter reduces the total number of texture bits [83]. Hence, TeXture bits (TX) is modelled as a function of frame activity (A) and quantization parameter (Q). Therefore, the total number of bits (T) of I, P, and B frame types can be represented as

$$T_\psi(A_\psi, Q_\psi) = TX_\psi(A_\psi, Q_\psi) + I_\psi \cdot MV_\psi, \tag{4.3}$$

where $\psi \in \{I, P, B\}$ is the frame type and I_ψ is an indicator function defined as

$$I_\psi = \begin{cases} 1, & \text{if } \psi \in \{P, B\} \\ 0, & \text{if } \psi \in \{I\} \end{cases} . \tag{4.4}$$

Note that the activity of I frame (A_I) is calculated using unquantized, DCT transformed I image whereas the activities of P and B frames (A_p and A_B) are calculated using unquantized, motion-compensated DCT transformed P and B images. Assumptions made are that no shape coding is used for MPEG4 video sequence and there is only one rectangular object i.e. one VOP.

In the remaining sections, the terms texture bits and texture size are used interchangeably. The terms motion vector bits and motion vector size are also used interchangeably.

4.2.3 Measurement Methods and Video Sequences

Measurements are made during the encoding process to estimate the frame activity (A), output texture size (TX) and motion vector size (MV). Fig. 4.1 shows

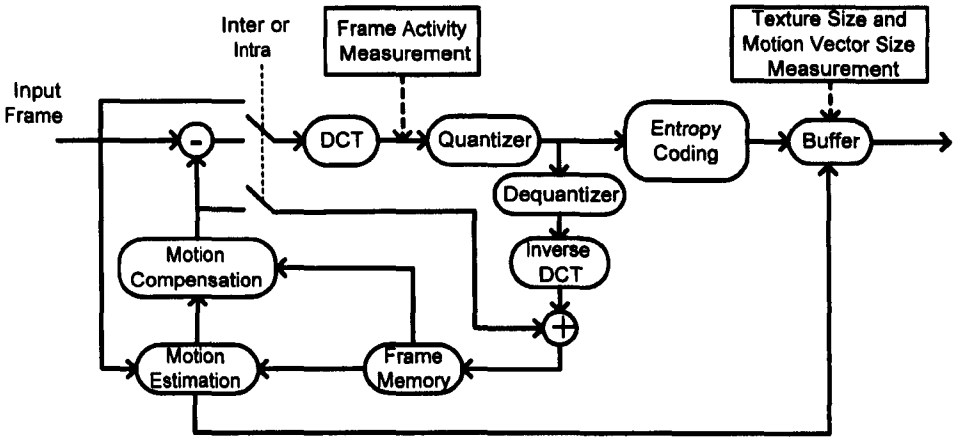


Figure 4.1: Frame activity, texture size, and motion vector size measurement points

the positions where the measurements are made during the encoding process. Each measurement results in a single entry consisting of

$$M = (\hat{A}_\Psi, \hat{T}X_\Psi, \hat{M}V_\Psi). \tag{4.5}$$

$\hat{M}V$ is replaced with zero for I frames since it does not have any motion vectors. The quantization parameters of I, P and B frames are set to be equal (i.e. $Q_I = Q_P = Q_B$) before the encoding process. The video is then encoded using MPEG4 codec and measurements are made for Eq. (4.5) to generate frame size traces. This procedure is carried out for typical MPEG4 quantization parameter range of $[1, 31]$ thus resulting in 31 traces where each trace is associated with a particular quantization parameter.

Video sequences used have a resolution of 176 × 144 (QCIF) since the study is intended for mobile multimedia communications where the display monitor is typically small. In the MPEG standard, the video sequence is typically encoded in a particular Group of Picture (GOP) pattern consisting of I, P, and B frame types. The GOP pattern used in this chapter is IBBPBBPBBPBB.

4.3 Frame Size Modelling

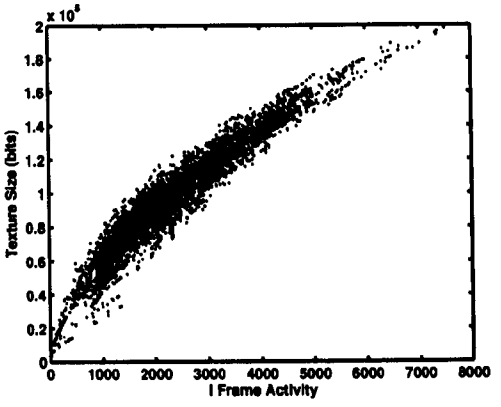
In this section, the techniques for mapping frame activity (A) to texture size (TX) for I, P, and B frame types are discussed. Then the marginal distributions of I, P, and B frame activities are fitted to suitable mathematical distribution. The marginal distributions of motion vectors of P and B frame types are also fitted to suitable mathematical distributions. Finally, the overall I, P and B frame size models is presented.

4.3.1 I, P and B Texture Size Modelling

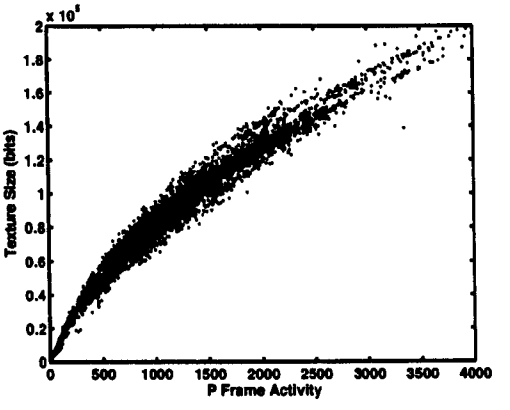
Frame Activity to Texture Size Mapping

The technique for mapping frame activity to texture size for a given quantization parameter is presented in this section. Scatter plots of frame activity (A) and the corresponding texture size (TX) of I, P and B frames for different quantization parameters of one are shown in Fig. 4.2. As illustrated in the figure, there is a strong correlation between frame activity and texture size. For curve fitting purposes, the activity range from zero to the largest calculated frame activity is divided into bins of size 50. All the values that fall within a particular bin are summed and averaged to obtained the mean value. The resulting curve is smoother and it facilitates the curve fitting process. The procedures are performed separately for I, P, and B frame types. It is found that this family of curves can be modelled accurately using quadratic functions for the activity range from zero to the largest calculated frame activity. Examples of fitted curves for the quantization parameter 5, 15 and 25 are plotted in Fig. 4.3. The quadratic function is defined as

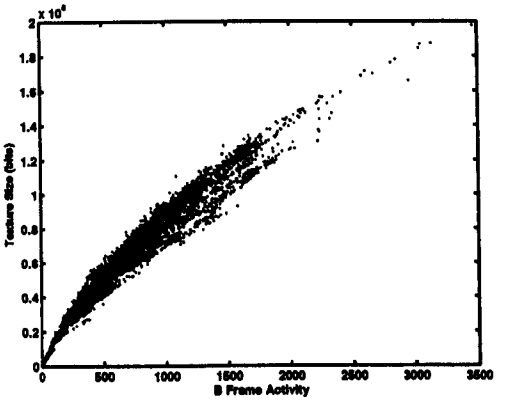
$$Y = C_0 \cdot X^2 + C_1 \cdot X + C_2. \tag{4.6}$$



(a) I frame

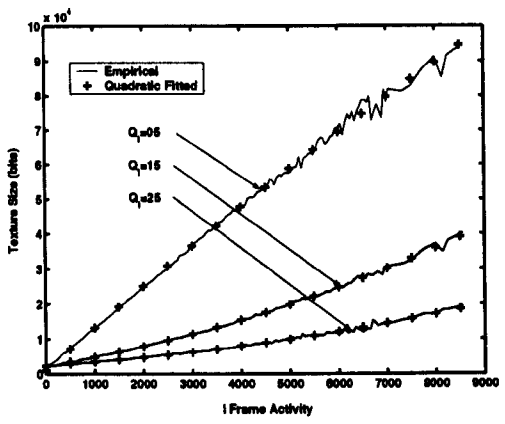


(b) P frame

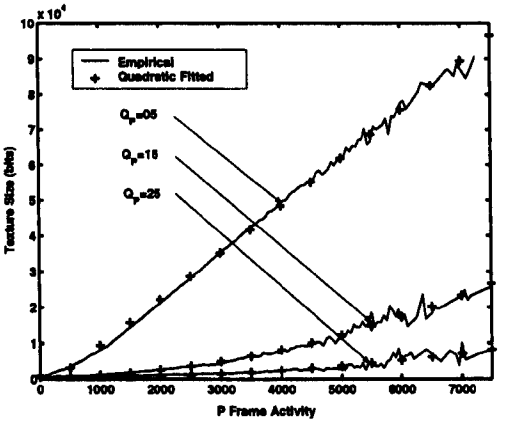


(c) B frame

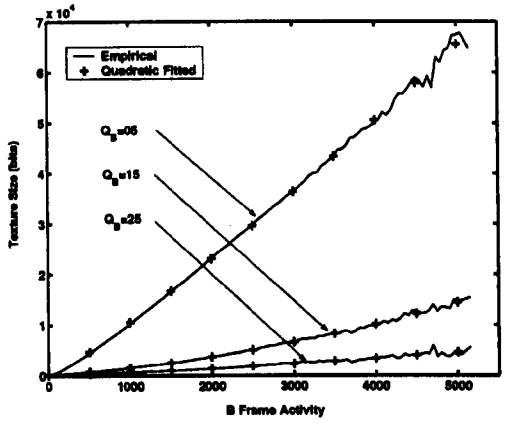
Figure 4.2: Scatter plot of frame activity against texture size for quantization parameter 1



(a) I frame



(b) P frame



(c) B frame

Figure 4.3: Quadratic function fitting of frame activity vs texture size curves for quantization parameter 5, 15, and 25.

Least square fitting is used to estimate (C_0, C_1, C_2) of the quadratic function. Each set of (C_0, C_1, C_2), together with its quantization parameter Q, is stored in a table as (Q, C_0, C_1, C_2) where Q also indicated the table index in this case. (Q, C_0, C_1, C_2) for I, P and B frames are calculated and stored in separate I, P and B Quadratic Coefficients Tables ($QCT^{(\Psi)}$). Using Eq. (4.6), the mapping of frame activity (A) to texture size (TX) for a given quantization parameter (Q) can be represented as

$$TX_\psi(A_\psi, Q_\psi) = C_0^{(\psi)}(Q_\psi) \cdot A_\psi^2 + C_1^{(\psi)}(Q_\psi) \cdot A_\psi + C_2^{(\psi)}(Q_\psi), \qquad (4.7)$$

where $\psi \in \{I, P, B\}$ is the frame type.

I, P and B Frame Activity Marginal Distribution Modelling

Cumulative Distribution Functions (CDF) of I, P, and B frame activity (i.e. A_I, A_P, and A_B) are plotted in Fig. 4.4. As shown in the figure, the Gamma distribution is found to fit the empirical distributions of I, P and B frame activity closely. The definition of Gamma distribution and its generation are discussed in Appendix B.1.

4.3.2 Motion Vector Size Marginal Distribution Modelling

With reference to Fig. 4.1, motion vectors of P and B frames are generated during the motion estimation process [83]. Once the motion estimation is completed, the total number of bits allocated for motion vectors is fixed. Motion vector size is independent of the quantization parameter and its marginal distribution can be assumed to be invariant over the whole quantization range $Q_\psi \in [1, 31]$. Hence the marginal distribution of P and B motion vector sizes ($\hat{MV}_P$, $\hat{MV}_B$) need to be fitted only once. As illustrated in Fig. 4.5, the marginal distributions of P and B motion vector sizes match the Gamma distribution closely. The definition of Gamma distribution and its generation are discussed in Appendix B.1.

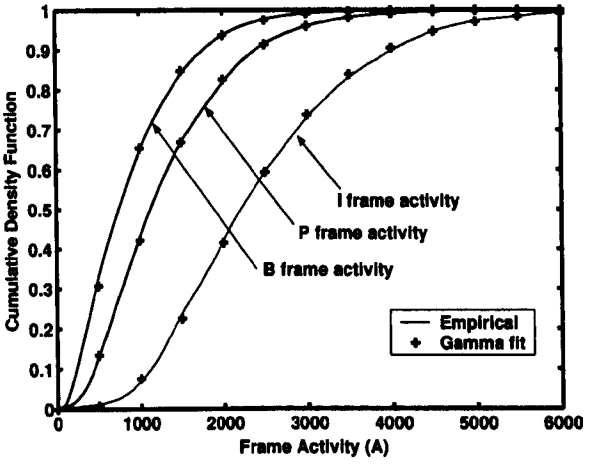


Figure 4.4: I, P, and B frame activity CDF fitted by Gamma distribution.

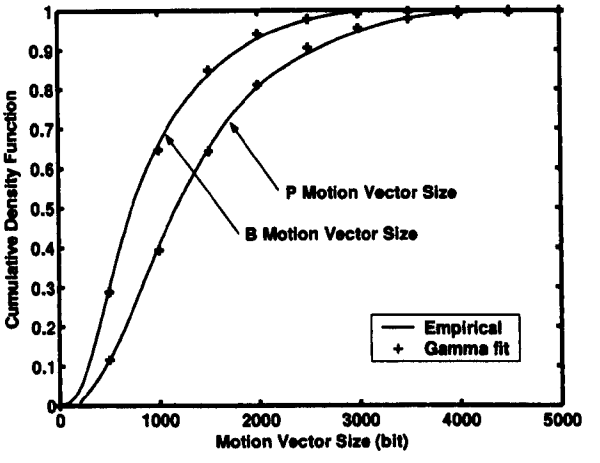


Figure 4.5: Motion vector size CDF fitted by Gamma distribution.

4.3.3 The Overall I, P, and B Frame Size Models

Based on previous discussions, the overall I, P, and B frame size models are presented in Fig. 4.6 and Fig. 4.7. Here the I frame size generation process is used as an example to explain the frame size generation process. First, a random Gaussian variable, X_1, is generated. Then X_1 is mapped to the A_I using a probability integral transform. Similar to Eq. (3.14), I frame activity can be calculated as

$$F_{A_I}(A_I) = F_G(X_1, \mu_1, \sigma_1) \tag{4.8}$$

$$A_I = F_{A_I}^{-1}(F_G(X_1, \mu_1, \sigma_1)), \tag{4.9}$$

where F_{A_I} and F_G are the CDF of A_I and the CDF of Gaussian variable. μ_1 and σ_1 are the mean and variance of X_1. Note that A_I is Gamma distributed as discussed in Section 4.3.1. Given the quantization parameter, Q_I, A_I can be mapped to I frame size, T_I, using Eq. (4.3) and $C_0^{(I)}$, $C_1^{(I)}$, and $C_2^{(I)}$ obtained from pre-calculated I frame Quadratic Coefficient Table, $QCT^{(I)}$. The Quadratic Coefficient Table is discussed in Section 4.3.1. P and B frame size generation processes are computed in a similar way.

4.4 Cross Correlation Modelling

The cross correlations between the estimated $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ are examined in this section. The cross correlations need to be considered for two reasons. Firstly, the output I, P and B frame sizes are highly correlated as discussed in Chapter 3.2. This is due to the cross correlation of underlying $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$.

Table 4.1 presents the cross correlation matrix for $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ where the highest correlation coefficient (excluding 1.0000 for the case of self cor-

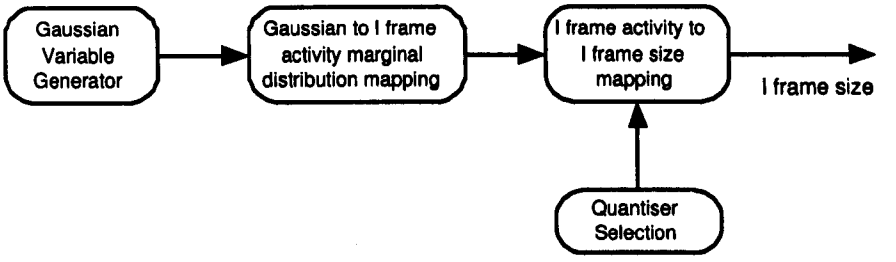


Figure 4.6: I frame size model.

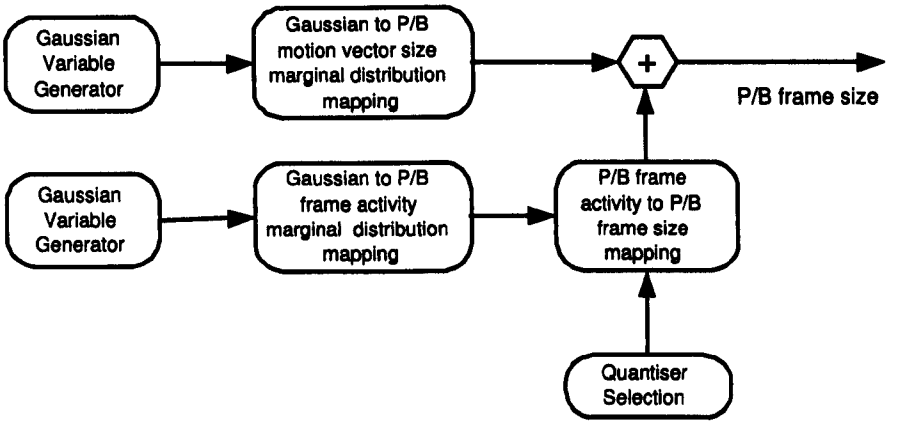


Figure 4.7: P and B frame size model.

	A_I	A_P	A_B	MV_P	MV_B
A_I	1.0000	0.8378	0.7569	0.0136	-0.1730
A_P	0.8378	1.0000	0.9192	0.4125	0.1418
A_B	0.7569	0.9192	1.0000	0.4567	0.2892
MV_P	0.0136	0.4125	0.4567	1.0000	0.7883
MV_B	-0.1730	0.1418	0.2892	0.7883	1.0000

Table 4.1: Correlation Matrix Between Model Parameters

relation) is 0.9192. Ignoring the cross correlation between $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ leads to underestimation of empirical traffic burstiness if the model is used as a traffic generator. This in turn underestimates the network queuing performance and subsequent simulations that depend on the generator will be inaccurate. Secondly, the P and B frame size marginal distributions obtained with Eq. (4.3) may not match the empirical marginal distribution if the texture size and motion vector size are generated independently. This is because the probability of a particular texture size coinciding with a particular motion vector size is not preserved. Accurate modelling of P and B frame sizes may still be achieved if the cross correlation between texture size and motion vector size are modelled. This issue is further examined in Section 4.7.2.

The cross correlations between $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ are measured using the cross correlation function in Eq. (3.1). The frame size traces generated with $Q_I = Q_P = Q_B$ are used for the calculation of correlation coefficients between $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$. The $\{M\}$ sequence from Eq. (4.5) is first parsed into separate I, P, and B frame size sequences. Note that in this work, all the data points for P and B frame types within a GOP ($I_1B_2B_3P_4B_5B_6P_7B_8B_9P_{10}B_{11}B_{12}$) are summed and averaged respectively to form a new sequence with GOP pattern $I_1P_{avg}B_{avg}$ before the parsing process. This is done so that the output I, P,

and B sequences have the same length and cross correlation coefficients between variables can be calculated. The calculated correlation coefficients are presented in Table 4.1.

The MultinoMial method (MM) discussed in Chapter 3.3 is used to model the cross correlation between (A_I, A_P, A_B, MV_P, MV_B). The MultinoMial method is used to generate correlated Gaussian vector $X = (X_1, ..., X_5)$ which will be mapped to (A_I, A_P, A_B, MV_P, MV_B) respectively. Similar to Chapter 3.3, a set of correlated variables can be obtained below

$$X_1^{(N)} = \frac{L_{11}Z_1}{\sqrt{L_{11}^2}} = Z_1 \tag{4.10}$$

$$X_2^{(N)} = \frac{L_{21}Z_1 + L_{22}Z_2}{\sqrt{L_{21}^2 + L_{22}^2}} \tag{4.11}$$

$$X_3^{(N)} = \frac{L_{31}Z_1 + L_{32}Z_2 + L_{33}Z_3}{\sqrt{L_{31}^2 + L_{32}^2 + L_{33}^2}} \tag{4.12}$$

$$X_4^{(N)} = \frac{L_{41}Z_1 + L_{42}Z_2 + L_{43}Z_3 + L_{44}Z_4}{\sqrt{L_{41}^2 + L_{42}^2 + L_{43}^2 + L_{44}^2}} \tag{4.13}$$

$$X_5^{(N)} = \frac{L_{51}Z_1 + L_{52}Z_2 + L_{53}Z_3 + L_{54}Z_4 + L_{55}Z_5}{\sqrt{L_{51}^2 + L_{52}^2 + L_{53}^2 + L_{54}^2 + L_{55}^2}}. \tag{4.14}$$

The individual element of vector $X^{(N)} = (X_1^{(N)}, ..., X_5^{(N)})$ is a zero mean and unity variance Gaussian variable and can be respectively mapped to (A_I, A_P, A_B, MV_P, MV_B) using probability integral transform as discussed in Section 4.3.3. The generated A_I, A_P, A_B, MV_P, MV_B have the same cross correlation behaviour to the estimated $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$.

4.5 Autocorrelation Modelling

The autocorrelation structure of traffic is modelled using the Spatial Renewal Process (SRP) discussed in Chapter 3.5.1. It is worth mentioning that other autocorrelation modelling techniques, e.g. wavelets, may be used and the model is

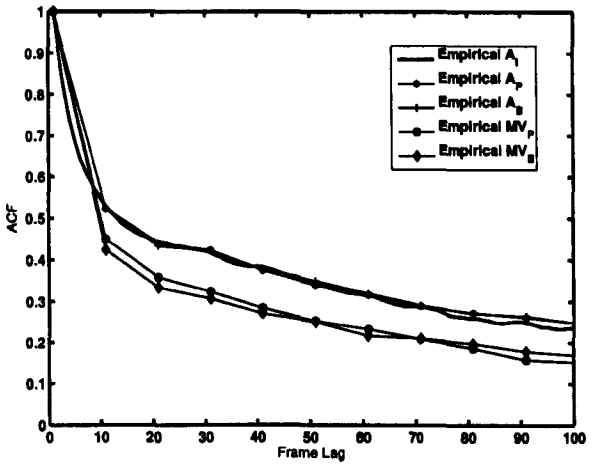


Figure 4.8: Autocorrelation plot for I, P and B frame activity and P and B motion vector size

not limited to SRP. Fig. 4.8 shows the autocorrelation of $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$. Note that the autocorrelation structures of $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ in this plot are calculated based on the modified GOP structure $I_1 P_{avg} B_{avg}$ discussed earlier in Section 4.4. It can be observed that all $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ exhibit similar autocorrelation structures. This is due to the fact that they are interdependent. In order to explain this clearly, suppose that there are two highly interdependent series, $\{V_1\}$ and $\{V_2\}$. As they are highly interdependent, the values of V_1 and V_2 would follow a similar trend or sample path i.e. both increasing or decreasing at the same time even though there is some randomness in each sequence. This in turn translates into similarity in autocorrelation structure. Using the fact that $\hat{A}_I$, $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$ are interdependent, only the autocorrelation of $\hat{A}_I$ needs to be modelled. Then the autocorrelation of A_P, A_B, MV_P, and MV_B will derive their autocorrelation from the generated A_I respectively by cross correlation modelling described in Section 4.4.

Fig. 4.9 shows the generation of $X_1^{(N)}(n)$-$X_5^{(N)}(n)$. As discussed in previous sec-

$Z_1 : Z_1(1), Z_1(2), Z_1(3), Z_1(4), Z_1(5), Z_1(6), Z_1(7), Z_1(8), Z_1(9), Z_1(10), Z_1(11), Z_1(12)$

$X_1 : X_1(1), X_1(2), X_1(3), X_1(4), X_1(5), X_1(6), X_1(7), X_1(8), X_1(9), X_1(10), X_1(11), X_1(12)$

$X_2 : X_2(1), X_2(2), X_2(3), X_2(4), X_2(5), X_2(6), X_2(7), X_2(8), X_2(9), X_2(10), X_2(11), X_2(12)$

$X_3 : X_3(1), X_3(2), X_3(3), X_3(4), X_3(5), X_3(6), X_3(7), X_3(8), X_3(9), X_3(10), X_3(11), X_3(12)$

$X_4 : X_4(1), X_4(2), X_4(3), X_4(4), X_4(5), X_4(6), X_4(7), X_4(8), X_4(9), X_4(10), X_4(11), X_4(12)$

$X_5 : X_5(1), X_5(2), X_5(3), X_5(4), X_5(5), X_5(6), X_5(7), X_5(8), X_5(9), X_5(10), X_5(11), X_5(12)$

Figure 4.9: Subsampled series $X_1(n)$,...,$X_5(n)$ for $n = 1, ..., 12$ before mapping to I, P, B frame activity and P, B motion vector size

tions, the MultinoMial method always generates 5-tuple Gaussian variables which are interdependent. Suppose that a GOP with IBBPBBPBBPBB pattern is to be generated. First generate $X_1^{(N)}(n)$-$X_5^{(N)}(n)$ for $n = 1, ..., 12$ using the MultinoMial method. For calculating I frame, $X_1^{(N)}(1)$ is mapped to A_I with probability integral transform while $X_2^{(N)}(1)$-$X_5^{(N)}(1)$ are discarded. A_I is then mapped to I frame using Eq. (4.3). The same steps are taken for generating P and B frame sizes. By examining Fig. 4.9 closely, it can be noted that the autocorrelation structure of subsampled $X_1^{(N)}$ should approximate $\hat{A}_I$. However, the original $X_1^{(N)}$ has 12 times as many values in one GOP when compared to the estimated $\hat{A}_I$. In order to generate $X_1^{(N)}$ with an appropriate autocorrelation structure, all the video frames are re-encoded as I frame and the frame activity (see Section 4.2.1) calculation is performed to obtain the background activity process, $\hat{A}$. Hence, $\hat{A}_I(k) = \hat{A}(12 \times (k-1) + 1)$ for $k = 1, ..., \frac{K}{12}$ where K is the total number of video frames. $X_1^{(N)}$ can now be generated using the autocorrelation structure of $\hat{A}$. When the generated $X_1^{(N)}$ series is subsampled as in Fig. 4.9 and mapped to A_I, it has an approximated autocorrelation structure of $\hat{A}_I$. It can also be observed that $R_{\hat{A}_I}(j) \approx R_{\hat{A}}(12 \times j)$ [87] where R is the calculated autocorrelation function. Thus the autocorrelation structure of A_I is modelled. As for the autocorrelation structure of $X_2^{(N)} - X_5^{(N)}$, they are derived from $X_1^{(N)}$ by cross correlation modelling in Eq. (4.11)-(4.14). When A_P, A_B, MV_P, and MV_B

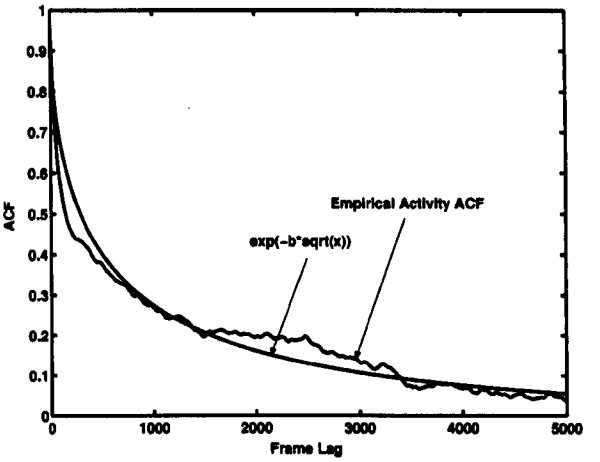


Figure 4.10: Background frame activity autocorrelation fitting

are calculated using the subsampled series of $X_2^{(N)} - X_5^{(N)}$, the autocorrelation structures produced are close to the autocorrelation of $\hat{A}_P$, $\hat{A}_B$, $\hat{MV}_P$, and $\hat{MV}_B$. With reference to Fig. 4.9, the summary of steps for the frame generation in a GOP is given below:

1. Use the SRP to model the autocorrelation structure of background activity process $\hat{A}$.

2. Generate a series of zero mean and unity variance Gaussian variables, Z_1, using SRP with the same autocorrelation structure to the background activity process $\hat{A}$.

3. Map Z_1 to $X_1^{(N)}$ using Eq. (4.10). Generate $Z_2, ..., Z_5$ as zero mean and unity variance random Gaussian variables. Map $Z_2, ..., Z_5$ to $X_2^{(N)}, ..., X_5^{(N)}$ using Eq. (4.11)-(4.14).

4. Subsample the series $X_1^{(N)}, ..., X_5^{(N)}$ as shown in Fig. 4.9 before mapping them to A_I, A_P, A_B, MV_P, and MV_B. Finally combine A_I, A_P, A_B, MV_P, and MV_B appropriately to obtain I, P, and B frames using Eq. (4.3).

Based on the previous discussion, it is necessary for the autocorrelation structure of background activity process $\hat{A}$ to be modelled accurately using SRP. The autocorrelation structure of background activity process $\hat{A}$ is calculated and shown in Fig. 4.10. It can be seen that the function $e^{-b\sqrt{k}}$ fits the autocorrelation of background activity process, $\rho(k)$, closely. Therefore,

$$\rho(k) = e^{-b\sqrt{k}}, \tag{4.15}$$

where k is the frame lag and b is the fitted coefficient. $\rho(k)$ is then used to calculate scene duration distribution $F_T(k)$ in Eq. (3.11). The marginal distribution of scene level used is assumed to be zero mean and unity variance Gaussian distribution.

4.6 Summary of GVTM for Traffic Generation

The proposed GVTM is illustrated in Fig. 4.11. The autocorrelation structure is modelled using SRP while cross correlations between frame sizes are modelled using the MultinoMial method. The frame size model is a tunable model that generates frame sizes for different quantization parameters. The building block "Adaptive Quantizer Select" is used to select appropriate quantization parameters. First, Z_1 is generated as a Gaussian distributed variable with zero mean and unity variance using SRP. The Z_1 has the same autocorrelation structure to the empirical source video sequence. Then the MultinoMial method that models the cross correlation is used to generate a zero mean, unity variance correlated Gaussian vector $X^{(N)} = (X_1^{(N)}, ..., X_5^{(N)})$ where $X_1^{(N)} = Z_1$. $X^{(N)} = (X_1^{(N)}, ..., X_5^{(N)})$ is transformed to (A_I, A_P, A_B, MV_P, MV_B) using probability integral transform. A_I, A_P, and A_B are further mapped to TX_I, TX_P, and TX_B using fitted quadratic function. TX_I is also the generated I frame size. P and B frame sizes are obtained by adding TX_P to MV_P and TX_B to MV_B. The generation of synthetic video traffic using the proposed GVTM is summarized below

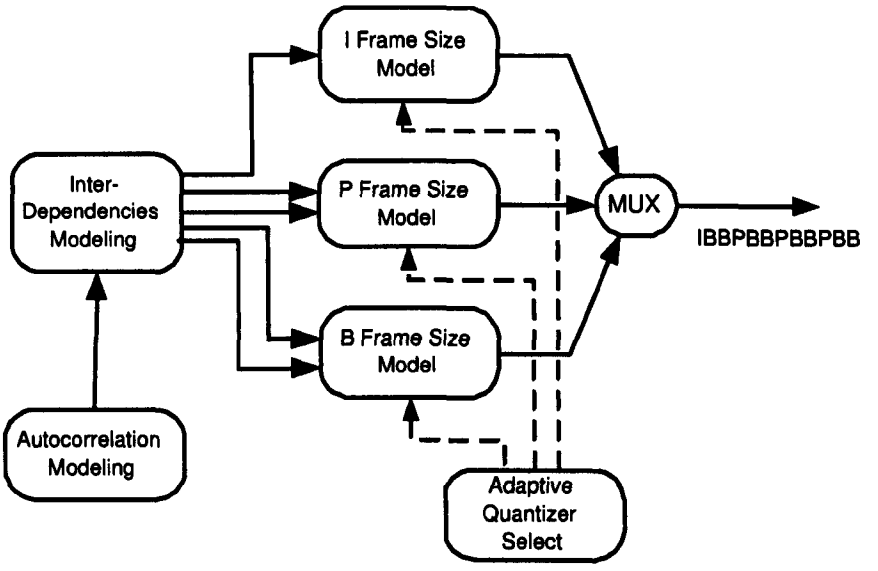


Figure 4.11: Block diagram of the proposed GVTM

1. Select a desired value for quantization parameter Q_ψ. Using Q_ψ as index to the table, load $C_0^{(\psi)}(Q_\psi)$, $C_1^{(\psi)}(Q_\psi)$ and $C_2^{(\psi)}(Q_\psi)$ from the pre-calculated Quadratic Coefficients Table, $QCT^{(\psi)}$, (see Section 4.3.1) for all $\psi \in \{I, P, B\}$.

2. Generate Z_1 as a zero mean, unity variance Gaussian distributed variable using SRP. Generate $Z_2, ..., Z_5$ randomly as zero mean, unity variance Gaussian distributed variables.

3. Calculate $X^{(N)} = (X_1^{(N)}, ..., X_5^{(N)})$ using $Z = (Z_1, ..., Z_5)$ and Eq. (4.10)-(4.14).

4. Determine the current frame type according to GOP pattern IBBPBBPBBPBB. Go to step 5 for I frame, step 6 for P frame and step 7 for B frame.

5. Map $X_1^{(N)}$ to A_I using Eq. (4.9). Map A_I to TX_I using $C_0^{(I)}(Q_I)$, $C_1^{(I)}(Q_I)$, $C_2^{(I)}(Q_I)$ and Eq. (4.7). In this case, TX_I is equal to output I frame size. Discard $X_2^{(N)}, ..., X_5^{(N)}$. Go to step 8 if all the required frames have been

generated. Otherwise go to step 1.

6. Map $X_2^{(N)}$ and $X_4^{(N)}$ to A_P and MV_P using the same technique in Eq. (4.9). Map A_P to TX_P using $C_0^{(P)}(Q_P)$, $C_1^{(P)}(Q_P)$, $C_2^{(P)}(Q_P)$ and Eq. (4.7). Sum TX_P and MV_P to obtain the output P frame size. Discard X_1^N, X_3^N, and X_5^N. Go to step 8 if all the required frames have been generated. Otherwise go to step 1.

7. Map $X_3^{(N)}$ and $X_5^{(N)}$ to A_B and MV_B using the same technique in Eq. (4.9). Map A_B to TX_B using $C_0^{(B)}(Q_B)$, $C_1^{(B)}(Q_B)$, $C_2^{(B)}(Q_B)$ and Eq. (4.7). Sum TX_B and MV_B to obtain the output P frame size. Discard $X_1^{(N)}$, $X_2^{(N)}$, $X_4^{(N)}$. Go to step 8 if all the required frames have been generated. Otherwise go to step 1.

8. Video traffic generation completed. Terminate.

4.7 Model Performance Evaluation

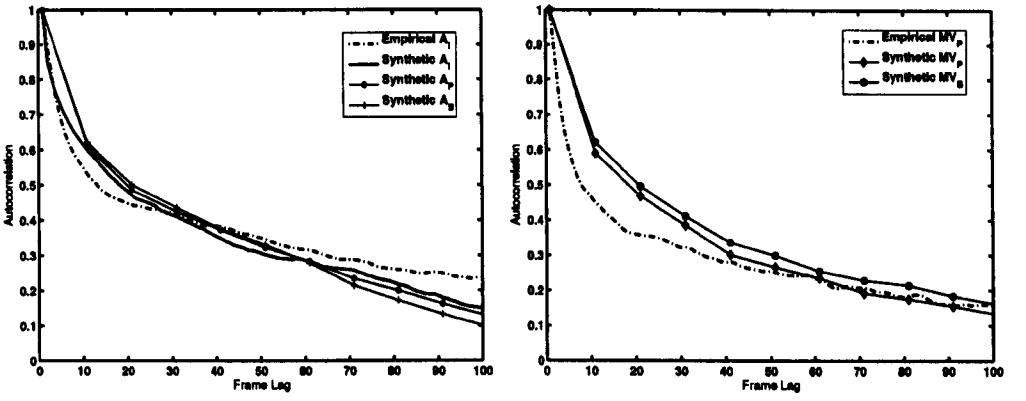
The results of GVTM validation are presented in this section. Three standard techniques from the literature are adopted for validating the proposed GVTM. They are the empirical autocorrelation matching, marginal distribution matching and empirical queuing performance prediction. GVTM is validated against different quantization parameter sets to show its applicability to a diverse range

Set	1	2	3	4	5	6
Q_I	05	15	25	05	10	10
Q_P	05	15	25	10	15	25
Q_B	05	15	25	15	15	25

Table 4.2: Quantization Parameter Set

of quantization parameter values. The quantization parameter set is listed in Table 4.2. Results shown are for the "The Lord of the Rings: Two Towers" video sequence with a total of 210936 frames.

4.7.1 Empirical Autocorrelation Matching



(a) Texture size (b) Motion vector size

Figure 4.12: Autocorrelation structure comparisons of synthetic texture size and motion vector size to their empirical counterpart

The synthetic autocorrelation structures of A_I, A_P, A_B, MV_P, and MV_B are shown in Fig. 4.12. The empirical autocorrelation structures of $\hat{A}_I$ and $\hat{MV}_P$ are also plotted as reference lines. It can be seen that the autocorrelation structures of GVTM generated traffic for A_I, A_P, A_B, MV_P, and MV_B are reasonably accurate when compared to their empirical counterparts.

4.7.2 Marginal Distribution Matching

Fig. 4.13 compares the CDF of GVTM generated frame sizes to that of empirical frame size for quantization parameters set 1 to set 6. The solid lines and the dashed lines corresponds to the empirical trace and the GVTM generated trace

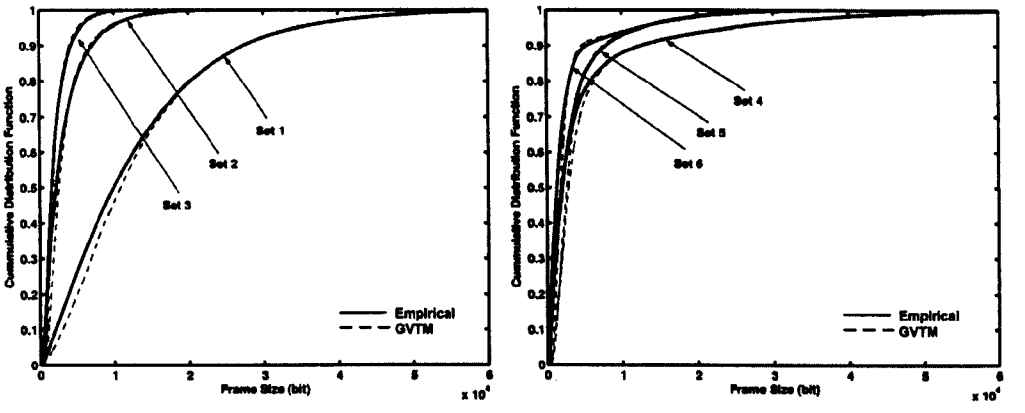
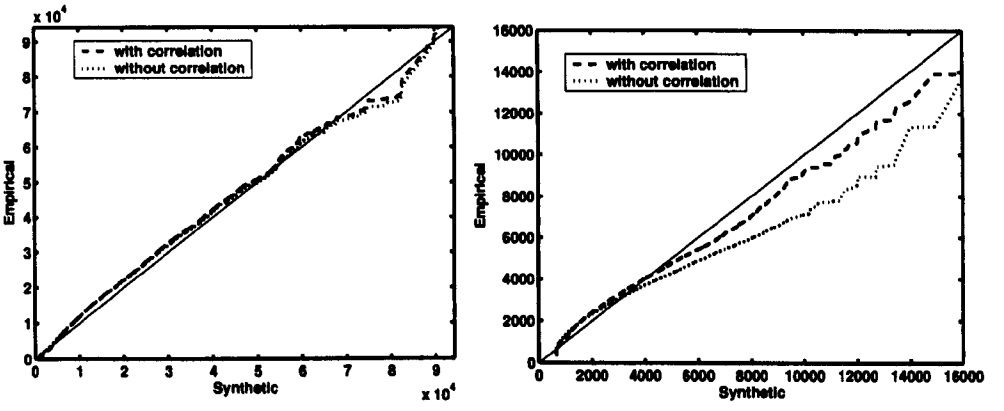


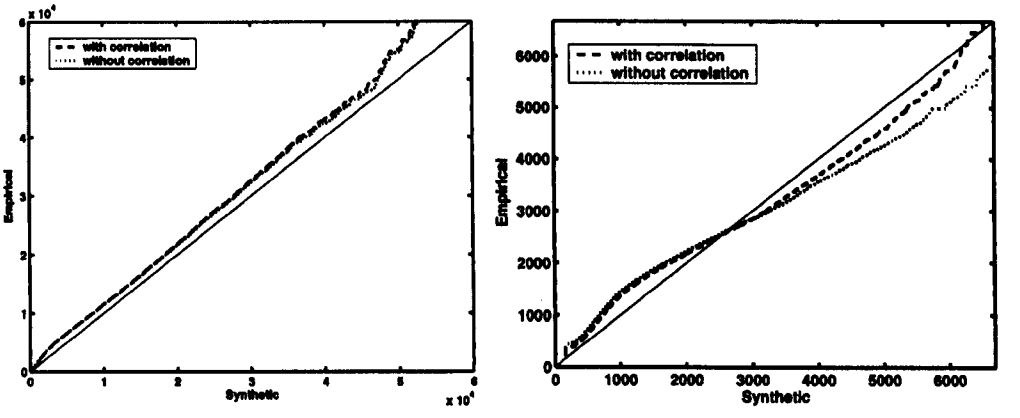
Figure 4.13: Comparisons of empirical frame size CDF to GVTM generated frame size CDF

respectively. It can be seen that the GVTM predicts the empirical frame size CDF closely for the whole video sequence. During the experiment, it is also verified that GVTM predicts the CDF of empirical frame sizes accurately for each I, P, and B frame type. The effect of ignoring the cross correlation between texture size and motion vector size is also examined. Fig. 4.14 illustrates the effect of ignoring the cross correlation between A_P and MV_p of P frame sizes using QQ plot for $Q_p = 5$ and $Q_p = 25$. When $Q_p = 5$, it can be seen that the CDF of generated P frame sizes without dependency modelling (i.e. TX_P and MV_P from Eq. (4.3) are generated independently without MM) are still closed to the CDF of empirical P frame sizes. This is because when the Q_P is small, the output texture size TX_P is much larger than the motion vector size MV_P. Therefore, MV_P is insignificant comparatively and accurate modelling of TX_P is sufficient to capture the CDF of empirical P frame sizes. However, when Q_P is large (i.e. image coarsely quantized), the texture size is comparable to motion vector size and the effect of ignoring the cross correlation between A_P and MV_p is more apparent. This can be clearly seen from Fig. 4.14. By modelling the cross correlation between the texture size and motion vector size, it has been verified that GVTM is capable of predicting the CDF of empirical frame sizes accurately



(a) Quantization parameter 5 (b) Quantization parameter 25

Figure 4.14: QQ plot for the synthesized P frame sizes with and without dependency modelling



(a) Quantization parameter 5 (b) Quantization parameter 25

Figure 4.15: QQ plot for the synthesized B frame sizes with and without dependency modelling

for all quantization parameters in the range $Q_\psi \in [1, 31]$. Similar results can be observed from Fig. 4.15 for B frames.

4.7.3 Queuing Performance Prediction

Existing Models for Comparison

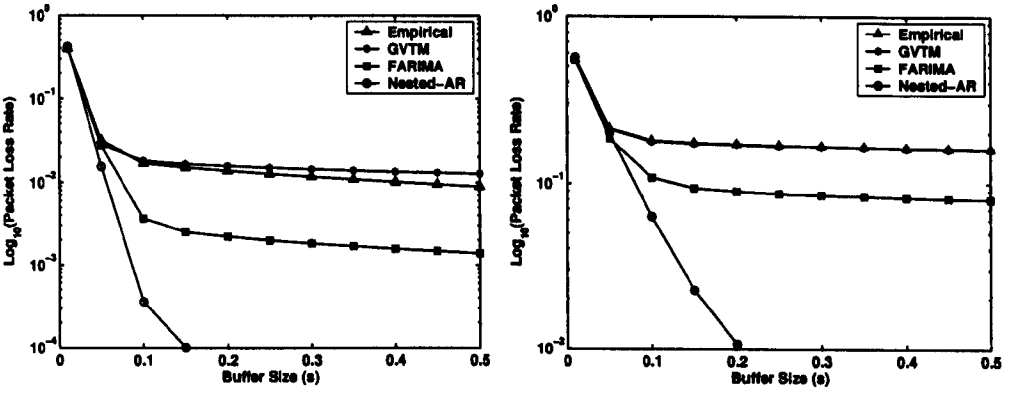
Two existing models are implemented for MPEG4 codec for comparison. They are the Nested AutoRegressive (Nested-AR) [31] and Fractional AutoRegressive Integrated Moving Average (FARIMA) model [36].

Simulation scenarios

The queuing system utilized for simulation is a finite buffer size First In First Out (FIFO) queue. Simulations are conducted using Network Simulator version two (NS2) [88]. The buffer sizes considered are in the range of 10 to 500 ms. The packet loss rates at bandwidth utilization $U = 40\%$, $U = 60\%$, and $U = 80\%$ are examined. Here, bandwidth utilization is defined as the ratio of mean video bandwidth to FIFO queue output bandwidth. For every bandwidth utilization, ten simulations are conducted for GVTM over the 10-500ms buffer size range using different random number generator seeds. The results (i.e. packet loss rate) from ten simulations are summed and averaged to obtain the mean value. The same simulation are repeated for the Nested-AR model and the FARIMA model using the same generator seed.

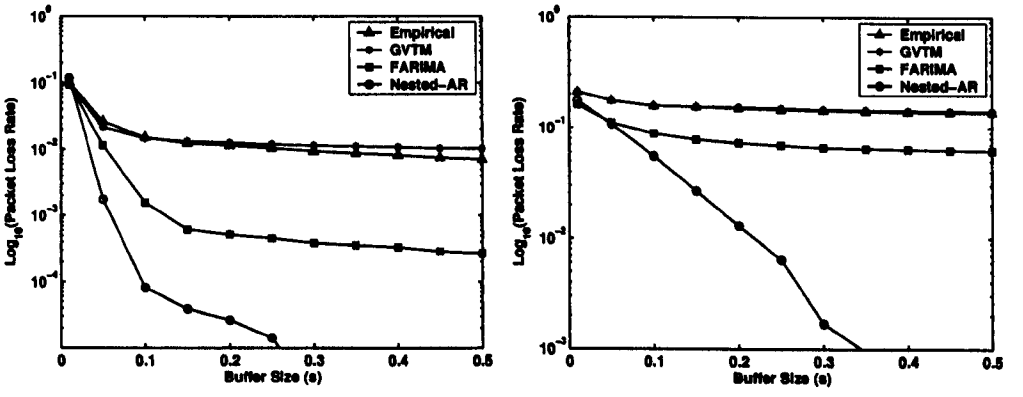
Packet Loss Rate Prediction

Simulations are performed for quantization parameter sets in Table 4.2. Fig. 4.16-4.21 compare the packet loss rate of empirical traffic to the model generated traffic



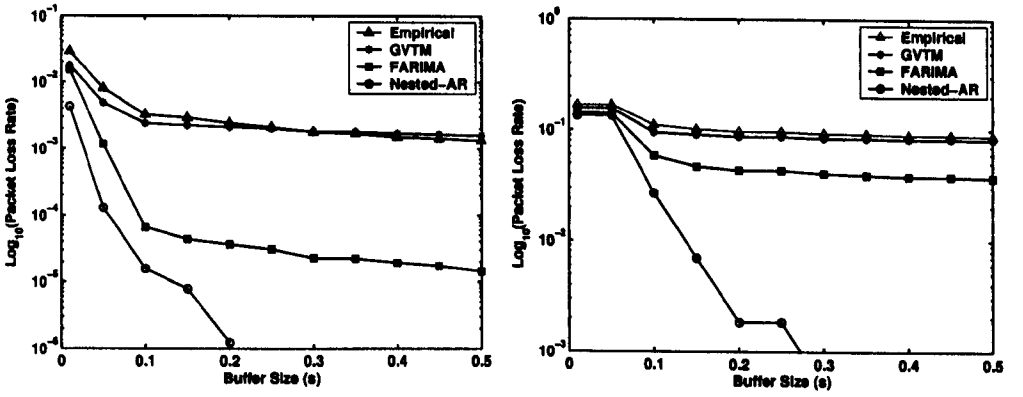
(a) Bandwidth utilization 40% (b) Bandwidth utilization 80%

Figure 4.16: Comparison of packet loss rate between empirical trace, GVTM, Nested-AR and FARIMA for quantization parameter Set 1



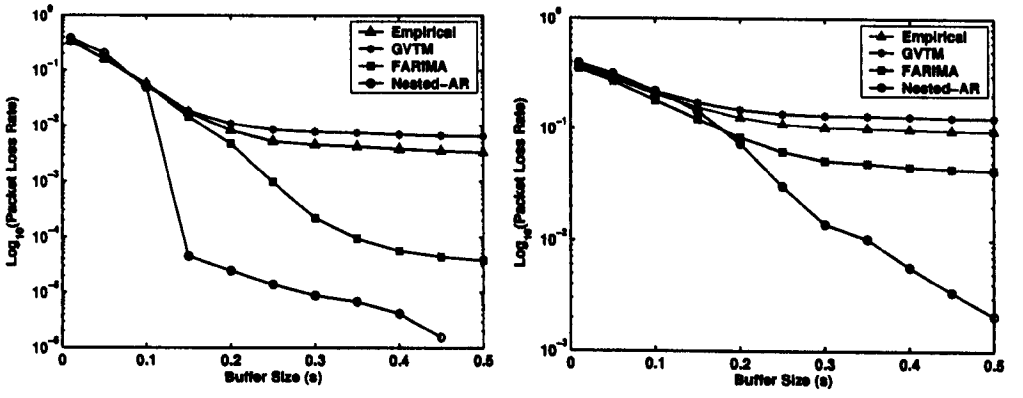
(a) Bandwidth utilization 40% (b) Bandwidth utilization 80%

Figure 4.17: Comparison of packet loss rate between empirical trace, GVTM, Nested-AR and FARIMA for quantization parameter Set 2



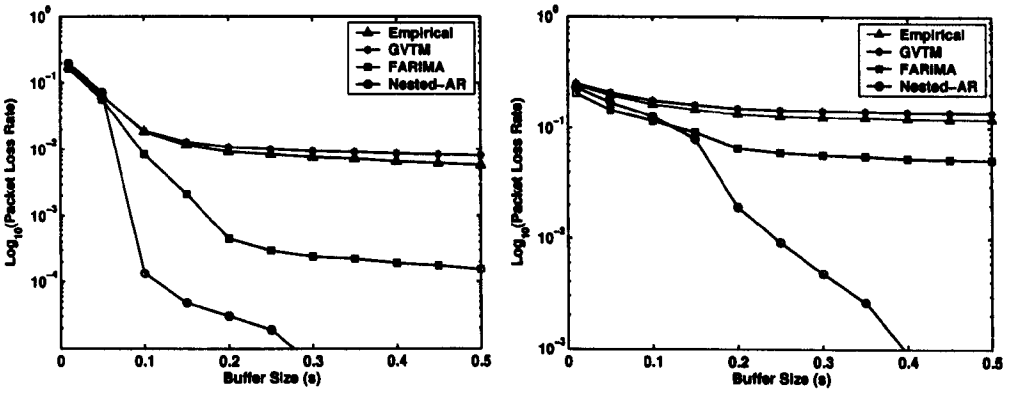
(a) Bandwidth utilization 40% (b) Bandwidth utilization 80%

Figure 4.18: Comparison of packet loss rate between empirical trace, GVTM, Nested-AR and FARIMA for quantization parameter Set 3



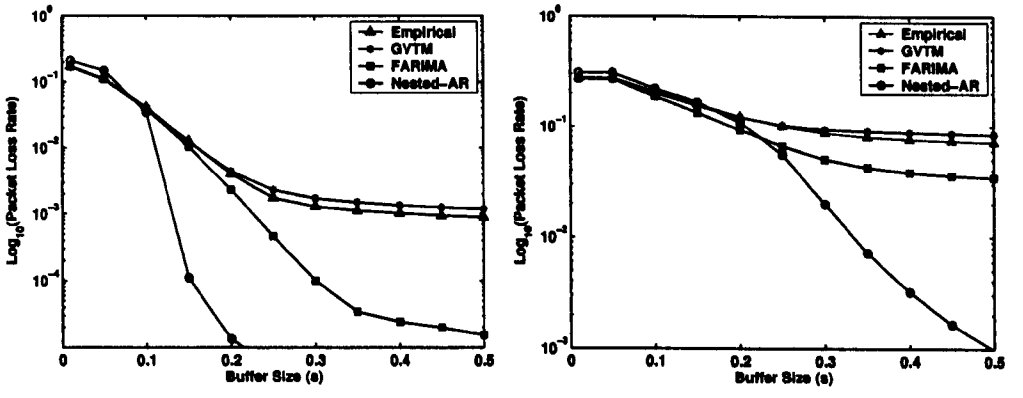
(a) Bandwidth utilization 40% (b) Bandwidth utilization 80%

Figure 4.19: Comparison of packet loss rate between empirical trace, GVTM, Nested-AR and FARIMA for quantization parameter Set 4



(a) Bandwidth utilization 40% (b) Bandwidth utilization 80%

Figure 4.20: Comparison of packet loss rate between empirical trace, GVTM, Nested-AR and FARIMA for quantization parameter Set 5



(a) Bandwidth utilization 40% (b) Bandwidth utilization 80%

Figure 4.21: Comparison of packet loss rate between empirical trace, GVTM, Nested-AR and FARIMA for quantization parameter Set 6

for quantization parameter sets in Table 4.2 for bandwidth utilizations of 40% and 80% respectively. Results for bandwidth 60% are not shown as they are similar. In all the figures, it can be observed that GVTM predicts the empirical packet loss rates accurately for all the buffer sizes and bandwidth utilizations considered. In contrast, Nested-AR and FARIMA models predict the empirical packet loss rate accurately for small buffer sizes, but the prediction accuracy degrades when the buffer size increases.

4.8 Conclusion

A Generalized Video Traffic Model (GVTM) is proposed in this chapter. The model provides a capability to generate output frame sizes for different quantization parameters in real time while considering inherent traffic characteristics such as autocorrelation structure and cross correlation between frame types. The developed GVTM can easily be configured to simulate VBR traffic for different quantization parameters as required in real time. This overcomes the time consuming process of model parameters re-estimation process that is commonly encountered in conventional VBR video traffic models when different sets of quantization parameters are required. Furthermore, GVTM has the advantage over the conventional model of allowing simulation of an adaptive source rate MPEG4 encoded video. This adaptive source rate capability is achieved in GVTM by modifying quantization parameters in real time so that the video source rate is adapted to the time-varying channel bandwidth. This is useful, for example, for the study of adaptive wireless video transmission where channel bandwidth fluctuates due to the physical environment conditions. The simulation results show that GVTM accurately captures the inherent characteristics of video traffic i.e. frame size marginal distribution, autocorrelation and cross correlation between frame types. Moreover, GVTM predicts the queuing performance of empirical

traffic with high accuracy for different quantization parameters, network bandwidth utilizations, and buffer sizes.

Chapter 5

Link Level and System Level Simulator for OFDM System

5.1 Introduction

This chapter presents the design and implementation of a link level simulator and a system level simulator for an Orthogonal Frequency Division Multiplexing (OFDM) system. First, the fundamentals and advantages of OFDM are presented. Secondly, implementation procedures of the link level simulator are presented. The link level simulator is used to generate block error rate performance curve for different Modulation and Coding Schemes (MCS). These performance curves can be utilized in system level simulators to model the error performance for a given Signal to Interference Ratio (SIR). Finally, implementation procedures of the system level simulator are discussed. Both simulators will be used in succeeding chapters for the study of multimedia resource allocation and scheduling.

5.2 OFDM Fundamentals

Orthogonal Frequency Division Multiplexing (OFDM) [89] has become one of the promising solutions for next generation broadband wireless air interface. In OFDM, a high rate stream is converted into multiple low rate streams for transmission. The symbol duration for any low rate stream is longer than the original high rate stream. Hence data transmission is more robust against Inter Symbol Interference (ISI) induced by multipath channels. This enables the multi-carrier OFDM system to transmit high bit rate multimedia streams with better error performance when compared to with the single carrier Code Division Multiple Access (CDMA) system. The following section describes the basic OFDM transmitter and receiver before proceeding to discuss other advantages of the OFDM based system.

5.2.1 OFDM Transmitter and Receiver

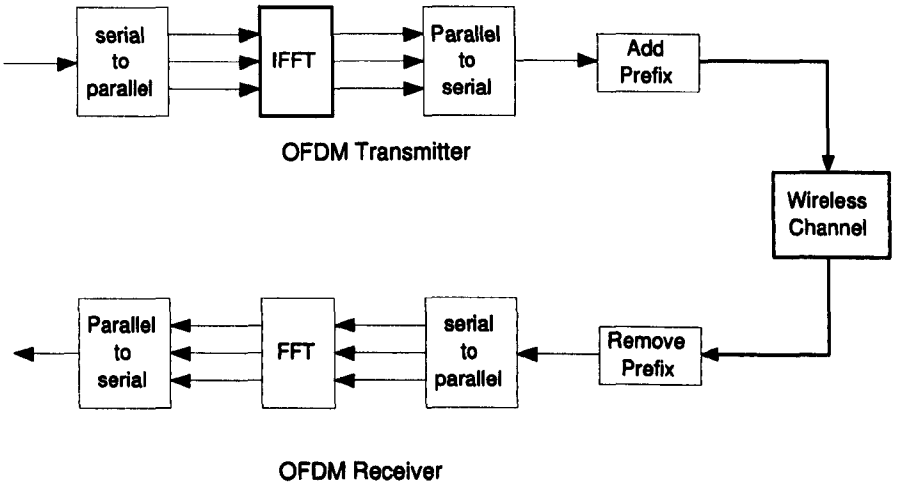


Figure 5.1: OFDM transmitter and receiver

Fig. 5.1 shows the basic structure of an OFDM transmitter and receiver: a stream with rate R bit/s is first converted to low rate stream with rate R/N bit/s where

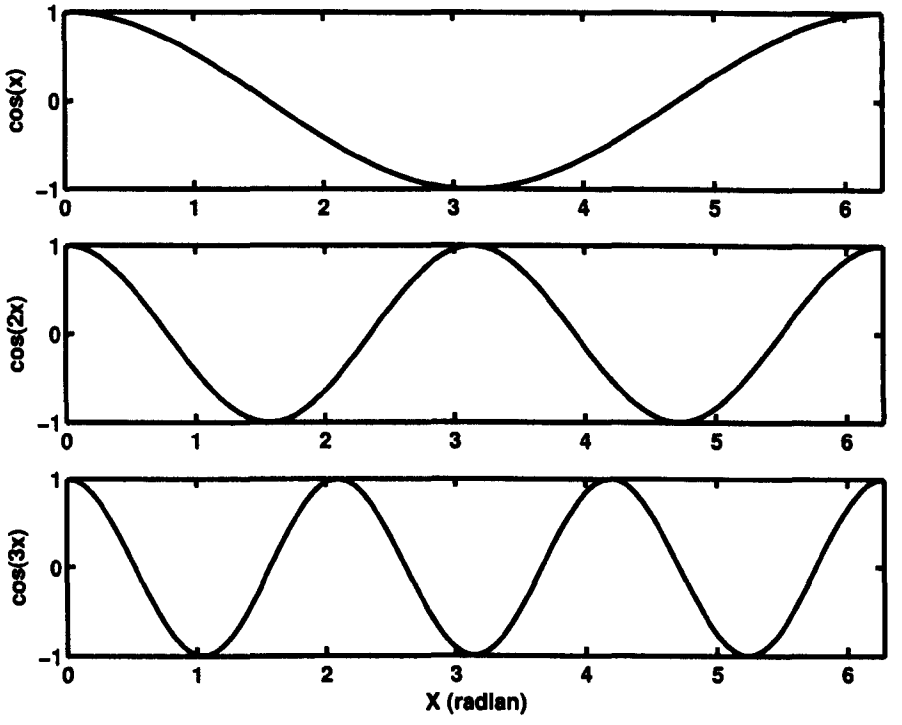


Figure 5.2: Subcarriers with frequency of 1, 2 and 3 Hz

N is the total number of subcarriers in the system. One bit from each low rate stream is grouped into an N-point data block. Inverse Fast Fourier Transform (IFFT) is then applied to transform the data block from frequency domain signal to time domain signal. Each point in the data block corresponds to a subcarrier with a certain frequency. In OFDM, subcarrier frequencies are chosen such that they are harmonics of some fundamental frequency f_o. Due to this property, subcarriers are orthogonal to each other. For example, Fig. 5.2 shows a subcarrier with fundamental frequency of 1Hz and its harmonics of 2Hz and 3Hz. It can be seen that 2 and 3 Hz harmonics have integral cycles of f_o. Due to this property, subcarriers are orthogonal to each other since

$$\int_0^T sin(2\pi k f_o) sin(2\pi l f_o) dt = 0, \qquad k \neq l. \tag{5.1}$$

Although subcarriers are orthogonal to each other, ISI may destroy the orthog-

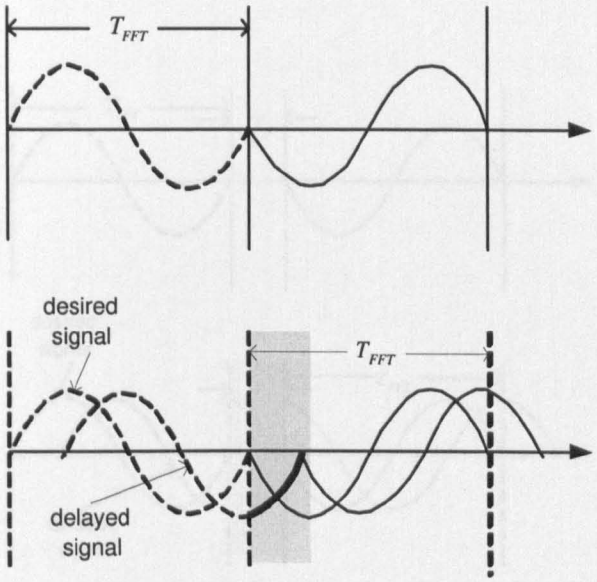


Figure 5.3: OFDM symbol without guard period. The figure is adapted from [89]

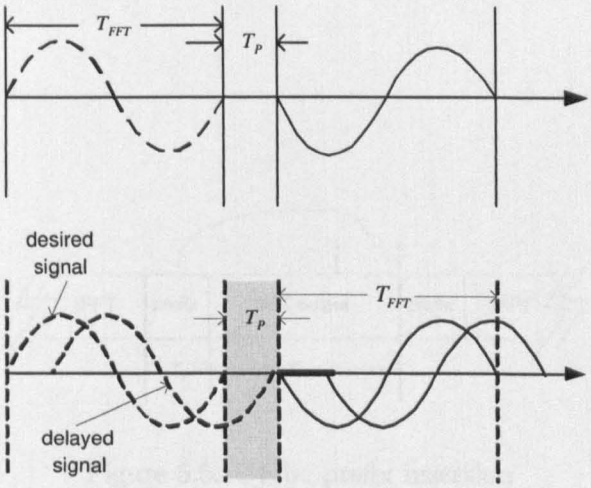


Figure 5.4: OFDM symbol with guard period. The figure is adapted from [89]

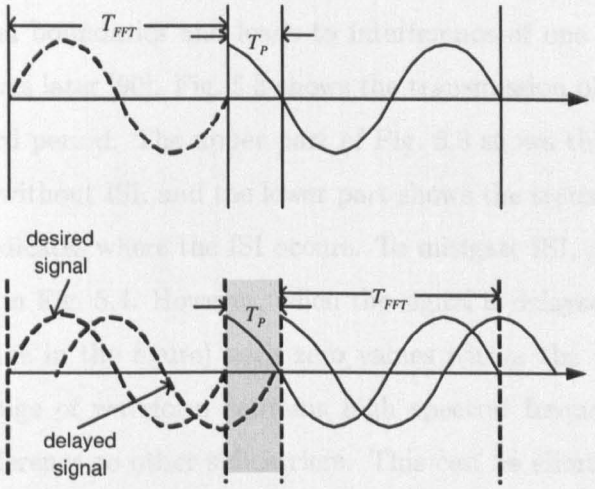


Figure 5.5: OFDM symbol with guard period and cyclicly extended prefix. The figure is adapted from [89]

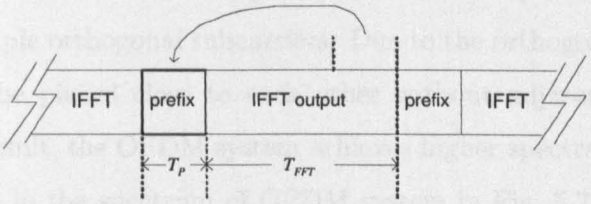


Figure 5.6: Cyclic prefix insertion

onality properties. This is because output radio signals travel through several propagation paths before reaching the receiver. These signals arrive at the receiver at different time instants as the distance on each path is different. This spreads the signal boundaries and leads to interference of one symbol with the symbol that arrives later [90]. Fig. 5.3 shows the transmission of OFDM symbols without the guard period. The upper part of Fig. 5.3 shows the transmission of OFDM symbols without ISI, and the lower part shows the transmission with ISI. The thick line indicates where the ISI occurs. To mitigate ISI, guard period, T_P, can be added as in Fig. 5.4. However, when the signal is delayed, there are some samples (thick line in the figure) with zero values within the window of T_{FFT}. The sudden change of waveform contains high spectral frequency components and causes interference to other subcarriers. This can be eliminated by cyclicly extending the waveform into the guard period as shown in Fig. 5.5. Cyclic extension is achieved by copying the tail part of FFT output to the guard period. The operation of cyclic extension is shown in Fig. 5.6.

5.2.2 Advantages of OFDM

High Spectral Efficiency

OFDM is similar to conventional Frequency Division Multiplexing (FDM), except that it uses multiple orthogonal subcarriers. Due to the orthogonality properties, subcarriers can be placed close to each other without adjacent band interference [90]. As a result, the OFDM system achieves higher spectral efficiency. This can be examined in the spectrum of OFDM system in Fig. 5.7. The sinc shape spectrum of each subcarrier is due to rectangular time windowing on output sine wave. The original line spectrum (sine wave) is convolved with the sinc shape spectrum (rectangle time window) in the frequency domain. It can be seen that the peak of each subcarrier corresponds to null of other subcarriers and interfer-

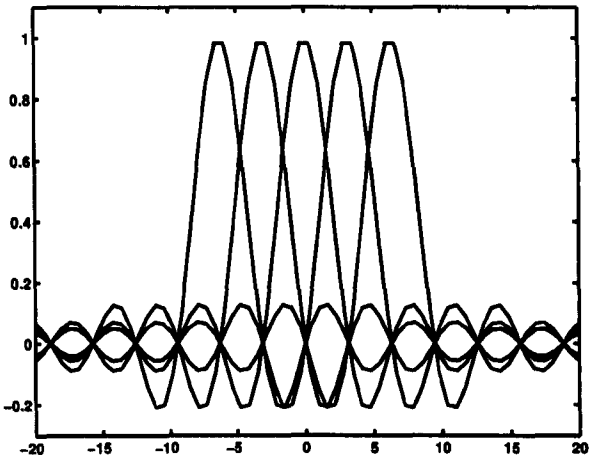


Figure 5.7: Spectrum of subcarriers

ence is avoided. This is an advantage over the conventional FDM system which requires frequency guard band to prevent adjacent bands interference.

It is also possible to avoid transmitting on deeply faded channel by dynamic subcarrier allocation [61] [91]. Fig. 5.8 shows an example of multipath channel gain for wireless environment. With proper subcarrier assignment, the deep channel fading at about subcarrier index 100 and 180 can easily be avoided by the user. By avoiding deeply faded subcarriers, the system is also more power efficient and a higher modulation level can be used to increase the spectral efficiency. The deeply faded subcarrier can still be allocated to another user since channel fading is uncorrelated between different users.

Efficient Multiple Access Scheme With Fine Granularity

OFDM divides the frequency spectrum into multiple orthogonal subcarriers. Each subcarrier or group of subcarriers can be allocated to a particular user. Hence, OFDM can be used as a multiple access scheme which is referred to as Orthogonal Division Frequency Multiple Access (OFDMA). OFDMA has the advantage of

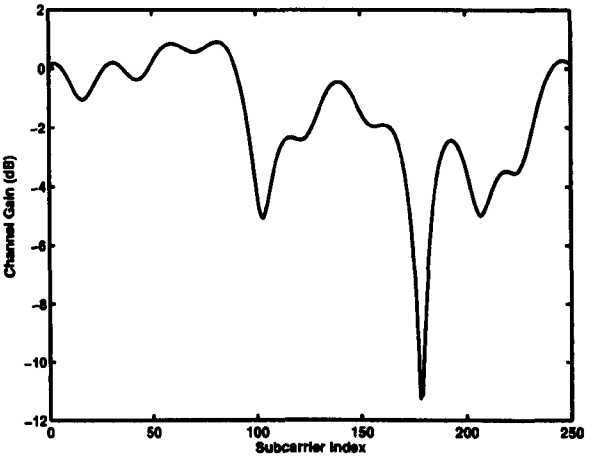


Figure 5.8: Wireless channel gain

fine granularity resource allocation where arbitrary subcarriers can be allocated. Thus, OFDMA can be more easily scaled to support heterogeneous application with different bit rate requirements. This avoids the inefficiencies of Time Division Multiple Access (TDMA) where the whole bandwidth spectrum is allocated to a single user even though the bit rate requirement of the user is low.

Furthermore, OFDMA can be combined with frequency hopping to improve the frequency diversity and interference diversity (e.g. Flash-OFDM [92]). Frequency Hopped OFDMA (FH-OFDMA) allows frequency reuse of one and facilitates radio network deployment without the costly and time consuming frequency planning process [93]. When compared with CDMA, FH-OFDMA avoids the intra-cell interference and is more robust to wideband fading. Power control inaccuracy in an OFDM based system is less stringent when compared with a CDMA system which suffers from the near far effect.

Simple Equalizer

OFDM transforms the frequency selective fading channel into parallel flat fading channels. This happens when the signal bandwidth of each subcarrier is smaller than the coherence bandwidth of the channel. For this reason, equalization in OFDM based systems can be achieved by a simple one-tap equalizer in frequency domain. Symbol equalization can be undertaken by one of several methods: 1) Phase compensation, 2) Maximum ratio combiner, 3) Zero forcing, and 4) Minimum mean square error. The reader is referred to [93] for descriptions of each equalizer.

5.3 Link Level Simulator

The link level simulator is described in this section. Implementation of the simulator closely follows the 3GPP TR25.892 technical report [93]. The simulator uses the IT++ signal processing library [94]. The main building block of the link level simulator includes cyclic redundancy check, turbo codec, rate matching, interleaving, QAM modem, channel multiplexing, OFDM transmitter and receiver and wireless channels. The overall simulator is illustrated in Fig. 5.9. Each component of the link level simulator will be described in the sections that follow.

The OFDM parameter set used in the simulator is shown in Table 5.1. With referring to Table 5.1, the total number of subcarriers in the system is $4.485MHz/15Khz = 299$. Assuming that there are 40 adjacent subcarriers in one traffic channel, then the maximum number of traffic channels is 7 since $\lfloor 299/40 \rfloor = 7$. The total number of symbols in one traffic channel for 2ms TTI is $40 \times 27 = 1080$. The link level simulator developed is for one traffic channel but it can be extended to include more channels if required.

Parameter	Value
TTI	2 ms
FFT size	512
OFDM sampling rate	7.68×10^6 sample/s
Cyclic prefix	56 samples
Subcarrier spacing	15 kHz
OFDM symbols per TTI	27
OFDM symbol duration	$73.96\mu s$
OFDM bandwidth	4.485 MHz

Table 5.1: OFDM Parameter Set [93]

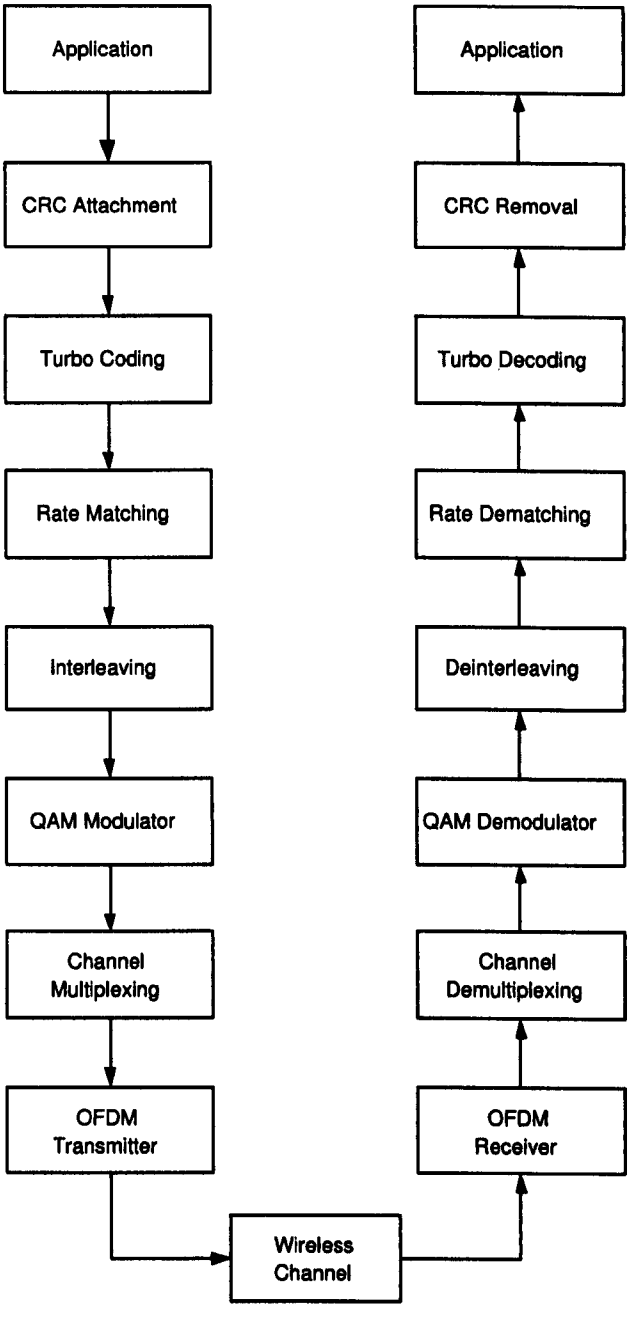


Figure 5.9: Link level simulator

5.3.1 Cyclic Redundancy Check

The Cyclic Redundancy Check (CRC) is used to ensure that the received data bits are error free. This is achieved by comparing the CRC checksum calculated before and after the transmission. The cyclic generator polynomial used is

$$G_{CRC}(D) = D^{24} + D^{23} + D^6 + D^5 + 1. \tag{5.2}$$

5.3.2 Turbo Codec

Turbo channel coding is used for Forward Error Correction (FEC). The scheme considered is a Parallel Concatenated Convolutional Code (PCCC) which consists of two 8 states convolutional encoders and an internal interleaver. The channel coding rate is 1/3. The polynomials of the convolutional encoders are

$$\begin{aligned} G_{Turbo,1}(D) &= D^3 + D^2 + 1 \\ G_{Turbo,2}(D) &= D^3 + D + 1 \end{aligned}. \tag{5.3}$$

Tail biting and internal interleaving procedures of turbo coding are discussed in [95]. A Log-MAX decoder [96] is used for decoding.

5.3.3 Rate Matching and Dematching

Rate matching is used to match the output bits of a turbo encoder to the total available bits in the traffic channel. This is achieved either by bit puncturing or repetition. Bit puncturing is performed on parity bits only. Zeros are inserted at the punctured position at the receiver before channel decoding. Bit repetition simply repeats some of the bits such that the total number of bits is equal to the total available bits in the traffic channel. Rate matching procedures are discussed in [95].

Rate matching is frequently used to control the data rate of a traffic channel according to the channel conditions. When channel conditions are favourable, rate matching can be used to reduce the amount of parity bits to allow more data bits. Conversely, rate matching allows more parity bits to be transmitted to provide greater channel protection when the channel conditions are becoming hostile.

5.3.4 Interleaver

Number of columns	Inter-column permutation pattern
30	0, 20, 10, 5, 15, 25, 3, 13, 23, 8, 18, 28, 1, 11, 21,6, 16, 26, 4, 14, 24, 19, 9, 29, 12, 2, 7, 22, 27, 17

Table 5.2: Inter-column permutation pattern for the interleaver [95]

An interleaver is used to spread the channel burst error over the time and frequency domain. In this way, a burst error appears as random noise and this improves the channel decoding performances. A block interleaver with the dimension of 72×30 is used in the simulator. The input bits are filled row by row into a matrix. If the number of input bits is less than 72×30, then dummy bits are used to pad the remaining elements of the matrix. Inter-column permutation is performed on the matrix to interleave the incoming bits. The permutation pattern is illustrated in Table 5.2. Finally, the output bits are read column by column from the interleaved matrix and the dummy bits are pruned from the output.

The interleaving operation is slightly different for 4QAM and 16QAM. Fig. 5.10 shows the interleaving operation for both 4QAM and 16QAM. For 4QAM, incom-

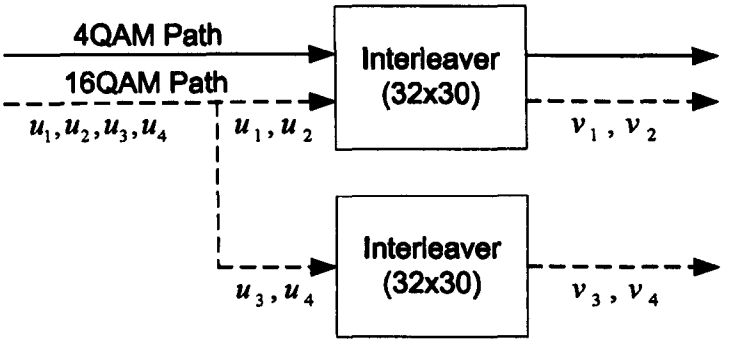


Figure 5.10: 4QAM and 16QAM interleaver

ing data bits follow the path shown as solid line into a block interleaver described previously. As for 16QAM, there are two similar interleavers in place. A group of 4 data bits are iteratively read from the input bit stream to fill the interleaver matrices. The first two bits go into the first interleaver and the two remaining bits go to the second interleaver. Once the interleaving operation is completed, output bits are collected in the same way as from the input bit stream. This is shown in Fig. 5.10 as dashed lines where u_1, u_2, u_3, u_4 are the incoming bits while v_1, v_2, v_3, v_4 are the output bits.

5.3.5 QAM Modulator and Demodulator

Quadrature Amplitude Modulation (QAM) uses two orthogonal frequency carriers to carry information. These carriers are called In-phase (I) and Quadrature (Q) carrier. Amplitudes of I and Q carriers can be used to convey data bits. This effectively maps to a unique location in two dimensional constellation diagram. For example, the constellation diagrams of 4QAM and 16QAM are shown in Fig. 5.11 and Fig. 5.12 respectively. Each point in the constellation diagram corresponds to a certain bit pattern. In 4QAM, each symbol can convey 2 data bits since there are $2^2 = 4$ constellation points. Similarly, 16QAM can convey 4 data bits in each symbol since $2^4 = 16$. The corresponding mapping of data

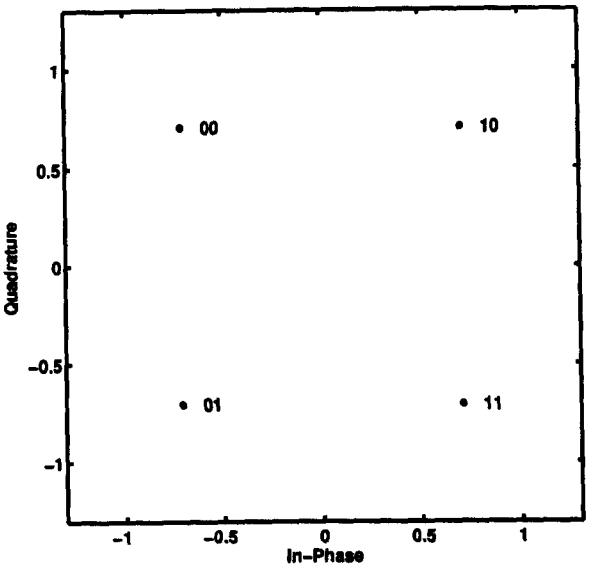


Figure 5.11: 4QAM constellation

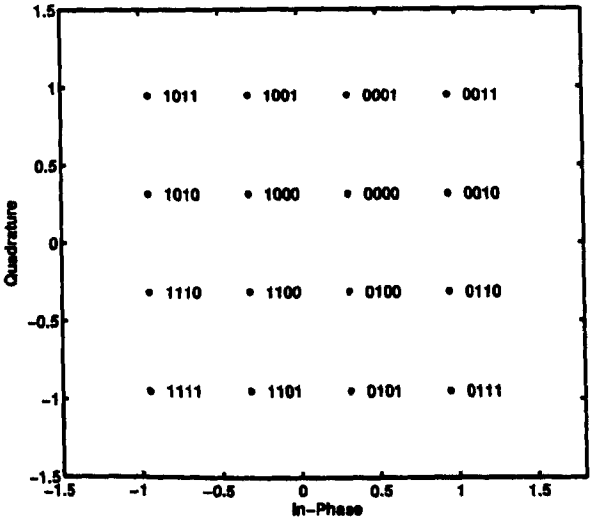


Figure 5.12: 16QAM constellation

Symbol	I	Q
00	0.7071	0.7071
01	0.7071	-0.7071
10	-0.7071	0.7071
11	-0.7071	-0.7071

Table 5.3: 4QAM modulation mapping

Symbol	I	Q	Symbol	I	Q
0000	0.3162	0.3162	1000	-0.3162	0.3162
0001	0.3162	0.9487	1001	-0.3162	0.9487
0010	0.9487	0.3162	1010	-0.9487	0.3162
0011	0.9487	0.9487	1011	-0.9487	0.9487
0100	0.3162	-0.3162	1100	-0.3162	-0.3162
0101	0.3162	-0.9487	1101	-0.3162	-0.9487
0110	0.9487	-0.3162	1110	-0.9487	-0.3162
0111	0.9487	-0.9487	1111	-0.9487	-0.9487

Table 5.4: 16QAM modulation mapping

bits to I and Q amplitudes for 4QAM and 16QAM are shown in Table 5.3 and Table 5.4. Both mapping tables are obtained from [97], but with normalization such that the average symbol energy is equal to one joule.

At the receiver, demodulation of QAM symbols involves the calculation of Log-Likelihood-Ratio (LLR) at each bit position [94]. The sign of LLR decides if a bit is 1 or 0 while the value of LLR indicates the certainty of the decision. A Turbo decoder can then utilize LLR as soft decision during the decoding process. The LLR is calculated using

$$\log\left(\frac{Pr(b_i=0|r)}{Pr(b_i=1|r)}\right) = \log\left(\frac{\sum_{s_i\in S_0}\exp\left(-\frac{|r-cs_i|}{N_o}\right)}{\sum_{s_i\in S_1}\exp\left(-\frac{|r-cs_i|}{N_o}\right)}\right), \tag{5.4}$$

where $s_i \in S_0$ denotes a set of QAM symbols with i^{th} bit equal to 0. r and c represent the received symbols and channel coefficient respectively. N_o is the single sided Additive White Gaussian Noise (AWGN) spectral density. In this case, the bit is decided as 0 if the calculated LLR is positive or 1 if the calculated LLR is negative.

5.3.6 Channel Multiplexing and Demultiplexing

Although the link level simulator only considers one traffic channel, it can easily be extended to more traffic channels. In this case, issues of channel multiplexing should be considered. Fig. 5.13 shows the time frequency resources of an OFDM system. Following on from previous discussions in the second paragraph of Section 5.3, G, g and n in Fig. 5.13 are equal to 280, 40 and 27 respectively. Each traffic channel consists of 1080 symbols and is labelled as Resource Unit (RU) in the figure. The output QAM symbols can be mapped onto any of the RUs during transmission. The resulting user data blocks are passed to the OFDM transmitter for transmission.

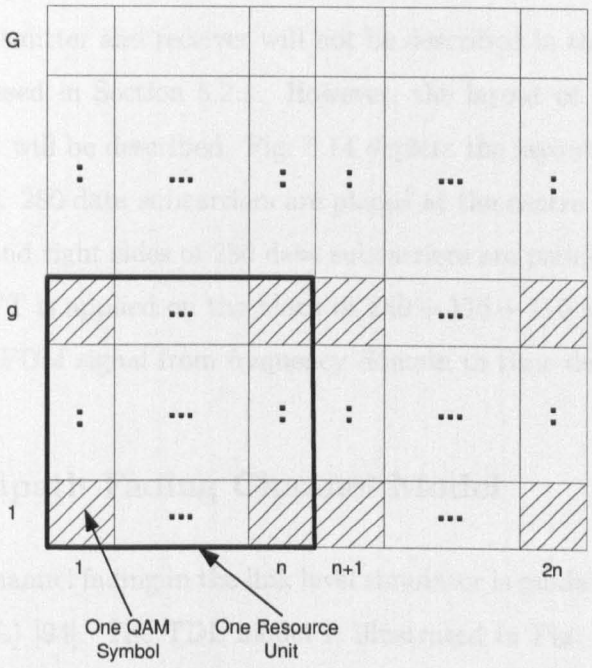


Figure 5.13: Time frequency resource in OFDM system

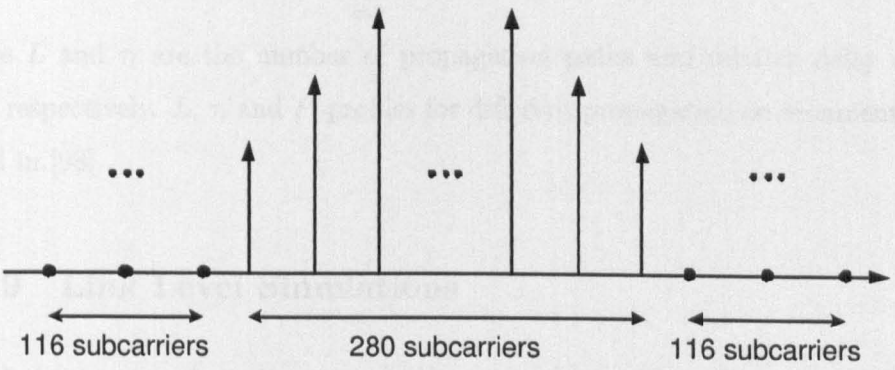


Figure 5.14: Subcarriers layout in 512-point IFFT data block

5.3.7 OFDM Transmitter and Receiver

The OFDM transmitter and receiver will not be described in this section as they have been discussed in Section 5.2.1. However, the layout of subcarriers in an IFFT data block will be described. Fig. 5.14 depicts the layout of subcarriers in IFFT data block. 280 data subcarriers are placed at the centre of the IFFT data block. The left and right sides of 280 data subcarriers are padded with 116 zeros respectively. IFFT is applied on the block of $280 + 116 + 116 = 512$ data points to convert the OFDM signal from frequency domain to time domain.

5.3.8 Multipath Fading Channel Model

The multipath channel fading in the link level simulator is modelled using Tapped Delay Line (TDL) [94]. The TDL model is illustrated in Fig. 5.15. P_0 to P_{L-1} represents the average power gain of each tap. $h_0(t)$ to $h_{L-1}(t)$ represents independent time-varying tap gain calculated using rayleigh distribution and classical doppler spectrum. The TDL is essentially a digital filter with impulse response of

$$h(t,\tau) = \sum_{l=0}^{L-1} h_l(t) \times P_l \times \delta(\tau - \tau_l), \tag{5.5}$$

where L and τ_l are the number of propagation paths and relative delay of l^{th} path respectively. L, τ_l and P_l profiles for different propagation environments are listed in [98].

5.3.9 Link Level Simulations

Simulations are conducted to estimate the transmission BLock Error Rate (BLER) for different Modulation and Coding Schemes (MCS) using the link level simulator described. Turbo coding with code rate of 1/3 is used in the simulation. The

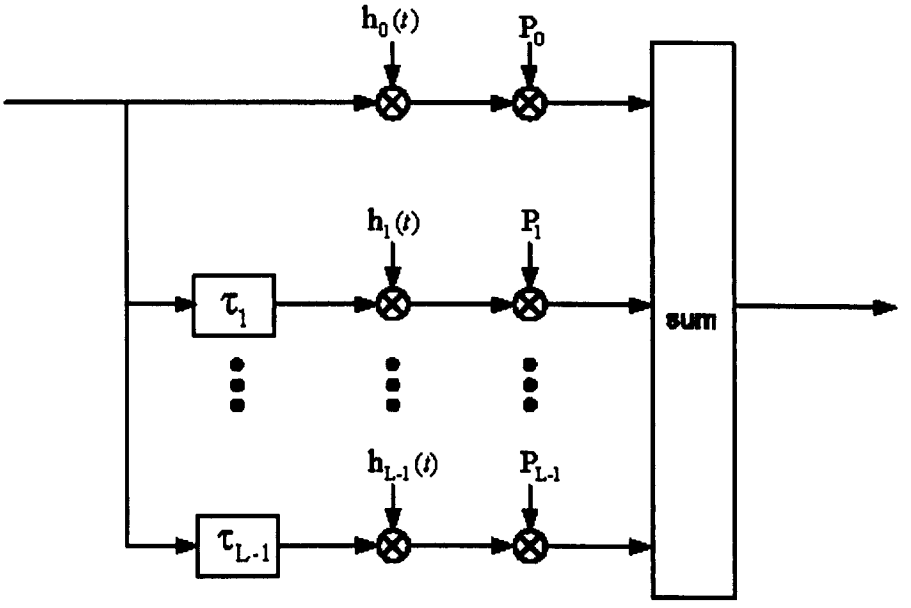


Figure 5.15: Tapped Delay Line (TDL) multipath channel model

block size, modulation scheme and the MCS code rate used for the simulation are illustrated in Table 5.5. The results of BLER performances for different MCS in the AWGN channel are illustrated in Fig. 5.16.

In a realistic transmission environment, frequency selective fading and other user interference may cause each QAM symbol within the traffic channel to experience different SIR. It is necessary to take this negative effect into account during the estimation of error performances. The Effective Exponential SIR Mapping (EESM) [99] method is adopted to take the variation of SIR in QAM symbols within the traffic channel into account. EESM maps QAM symbol SIRs into a single equivalent/effective AWGN SIR value. The Effective AWGN SIR value can then be used to estimate BLER from the simulated AWGN performance curves. The EESM function can be expressed as

$$SIR_{eff}(\beta) = -\beta \ln\left(\frac{1}{N_s}\sum_{k=1}^{N_s} exp\left(-\frac{\gamma_k}{\beta}\right)\right), \tag{5.6}$$

Scheme	Modulation	Block Size	Code Rate
MCS-1	4QAM	720	1/3
MCS-2	4QAM	1080	1/2
MCS-3	4QAM	1440	2/3
MCS-4	4QAM	1620	3/4
MCS-5	4QAM	1728	4/5
MCS-6	16QAM	1440	1/3
MCS-7	16QAM	2160	1/2
MCS-8	16QAM	2880	2/3
MCS-9	16QAM	3240	3/4
MCS-10	16QAM	3456	4/5

Table 5.5: Modulation and Coding Scheme (MCS) code rate.

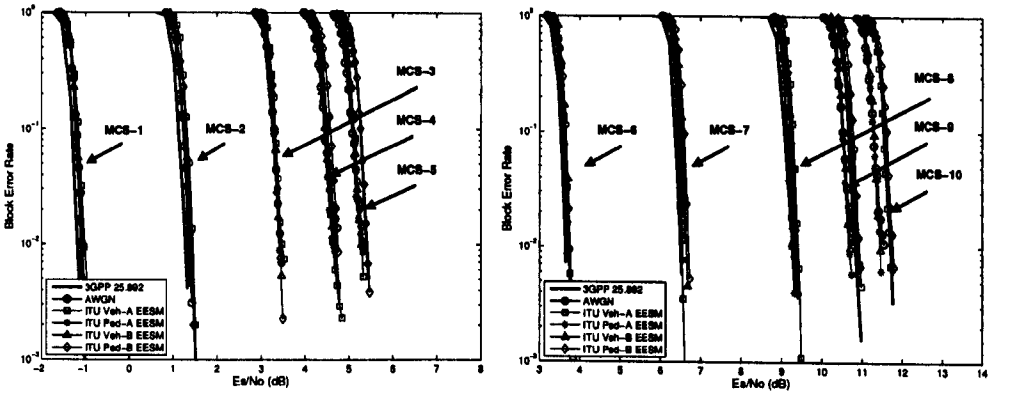


Figure 5.16: BLER performance for MCS-1 to MCS-10 for parameter in Table 5.1

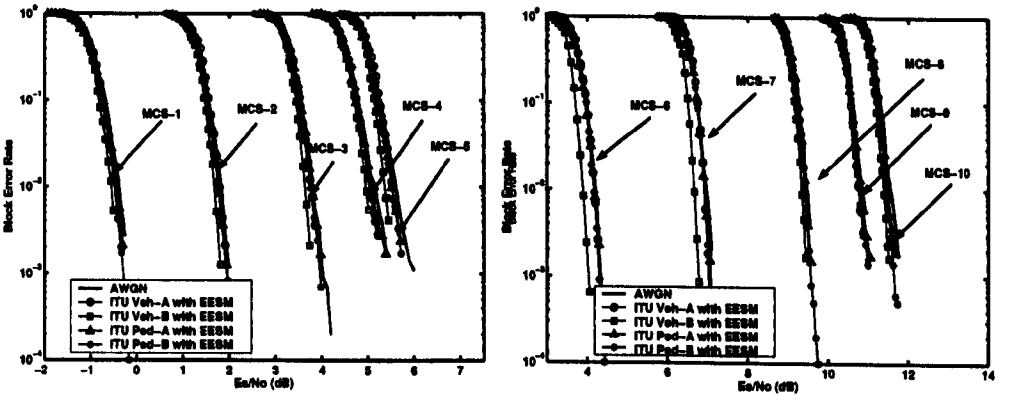


Figure 5.17: BLER performance for MCS-1 to MCS-10 for parameter set 2 [93]

Scheme	β	Scheme	β
MCS-1	1.24	MCS-6	2.88
MCS-2	1.44	MCS-7	4.31
MCS-3	1.61	MCS-8	6.59
MCS-4	1.73	MCS-9	7.72
MCS-5	1.69	MCS-10	7.77

Table 5.6: β for different MCS code rates

where N_s is the number of symbols in one traffic channel. γ_k is the SIR of a specific QAM symbol. β is a scalar parameter that must be estimated for every MCS. β for a given MCS can be estimated using minimum mean square error criteria

$$\beta = \arg\min_{\beta' \in [a,b]} \frac{1}{N_u} \sum_{m} [SIR_{awgn,m} - SIR_{eff,m}(\beta')]^2, \tag{5.7}$$

where a and b is range of values to be considered for estimation of β. N_u is the number of simulated BLER points. $SIR_{awgn,m}$ and $SIR_{eff,m}$ denote the SIR value of m^{th} data point for AWGN and realistic multipath channel respectively. Table 5.6 shows the estimated β for different MCS code rates using an ITU Vehicular A channel profile [98]. β for different channel profiles are found to be close in the simulation and also in [99]. Hence, β in Table 5.6 is sufficient for different channel profiles. Fig. 5.16 shows the examples of EESM mapped BLER curves for ITU Vehicular A, Vehicular B, Pedestrian A and Pedestrian B channel profiles from [98]. It can be seen that the EESM mapped BLER curves for different channel profiles matched to the AWGN BLER curves closely.

The simulator is configured to simulate OFDM link level performance for parameter set 2 in [93] for verification purposes. β values for different MCS code rate

are re-estimated for parameter set 2. The results of the simulation are shown in Fig. 5.17. The line with label "3GPP 25.892" shows the empirical results from [93]. The lines with label "AWGN" are the results for the link level simulator in AWGN environment. The lines with label "ITU Veh-A EESM", "ITU Ped-A EESM", "ITU Veh-B EESM" and "ITU Ped-B EESM" are the EESM mapped BLER curves for ITU Vehicular A, Pedestrian A, Vehicular B and Pedestrian B channel profiles. The results match the AWGN empirical results from [93] closely with maximum error of about 0.5dB.

5.4 System Level Simulator

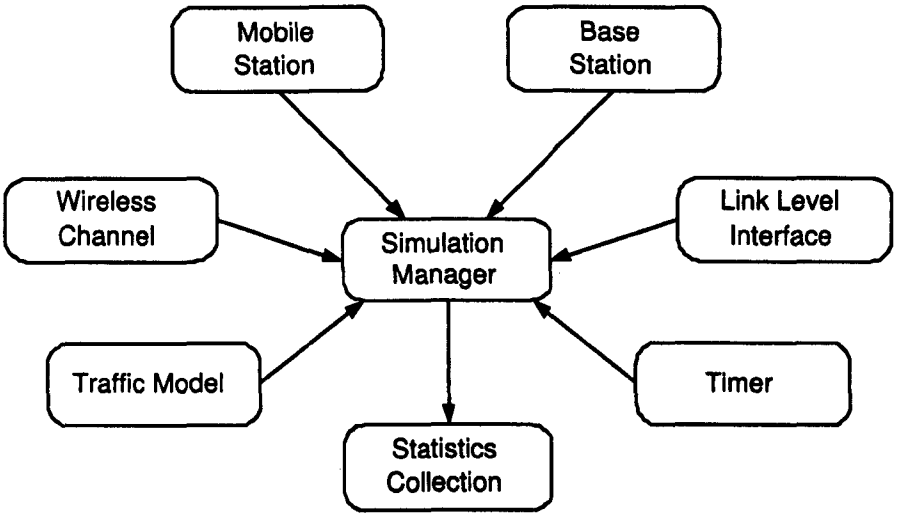


Figure 5.18: System level simulator model

The link level simulator aims at estimating the error performance of data transmission over wireless air interface. However, it provides no information on network performances of application. Network performances of interest may include average transmission bit rate, packet delay and delay violation probability etc. Hence a system level simulator is developed for studying the transmission of ap-

plication traffic over OFDM air interface. Fig. 5.18 shows the basic components of the developed system level simulator. Each component of the system level simulator will be discussed in the following sections.

5.4.1 Mobile Station

The mobile station component keeps track of all the movement and location coordinate of a user within a cell. The mobile is set to travel at some pre-defined speed at the beginning of the simulation. The direction of the mobile station is randomly selected at the beginning of a simulation, but is changed regularly after the mobile has travelled a certain distance. The probability of direction change is 0.2, with angle of direction change calculated from a uniform variable distributed in $[-45°, 45°]$.

5.4.2 Base Station

The basic function of a base station is to queue all incoming packets and dispatch them to the destination mobile station in a timely manner. Hence, all the resource allocation and packet scheduling algorithms are implemented in the base station component. It monitors queue and channel conditions of all the mobile stations, and dynamically schedules them for transmission. The Base Station then decides the transmission power and MCS on each traffic channel. Finally, packets are dispatched to destination mobile stations.

5.4.3 Wireless Channel Model

Wireless channel models considered in the system level simulator can be divided into three parts: 1) Path loss model, 2) Shadowing loss model, and 3) Multi-path fading model. Path loss and shadowing loss are calculated dynamically

during the simulation depending on the current location of the mobile station. As for the multipath channel model, it is computationally intensive to calculate the channel gain dynamically as it involves filtering operation of a tapped delay line. In fact, multipath channel modelling is the most computational demanding operation within the simulator. As an alternative, multipath channel gains are pre-calculated and stored in a set of traces. These traces can be repeatedly used during system level simulation and this reduces the overall computational complexity of the simulator. Path loss and shadowing loss models are described in the sections that follow, while the multipath fading channel model has been discussed in Section 5.3.8.

Path Loss Model

A vehicular test environment [98] path loss model is considered in the system level simulator. The path loss model is defined as

$$L = 40(1 - 4 \times 10^{-3}\Delta h_b)\log_{10}(R) - 18\log_{10}(\Delta h_b) + 21\log_{10}(f) + 80, \quad (5.8)$$

where L has a unit of dB. The path loss model has the following assumptions:

- Δh_b is the height difference between the base station antenna and average building rooftop height in metres. Δh_b is assumed to be 15m.
- R is the distance between mobile station and base station in km.
- f is the carrier frequency. f is assumed to be 2GHz

Shadowing Loss Model

The shadowing loss model used in the simulator is a correlated Log-Normal process [98]. The shadowing loss in dB, S, is calculated by

$$S(n) = C(d)S(n-1) + \sqrt{1 - [C(d)]^2} \times N(0, \rho), \quad (5.9)$$

where $N(0, \rho)$ is a random Gaussian variable with zero mean and standard deviation of ρ. $C(d)$ controls the autocorrelation of shadowing loss samples and is defined as

$$C(d) = 2^{-\frac{d}{d_{corr}}}, \tag{5.10}$$

where d and d_{corr} are the distance travelled and the decorrelation length respectively.

5.4.4 Link Level Interface

Link level simulations are usually computationally intensive and it would be infeasible to model link level error performance at the system level. Hence, a common technique used is to perform link level simulation separately and present the error performance to the system level simulator via a link level interface. One such interface is the Exponential Effective SIR Mapping (EESM) described in Section 5.3.9. In the system level simulator, the SIR of QAM symbols for a given traffic channel are mapped to an effective AWGN SIR value using the EESM. The corresponding block error rate can then be estimated from the simulated AWGN BLER curves.

5.4.5 Traffic Model

Traffic models are used to generate the amount of data traffic expected from an application to the destination mobile stations. This depends on the type of application and its characteristics such as average data rate and burstiness. Two basic applications are considered in this simulator, but more applications can be added if desired. The first application of interest is MPEG4 video traffic. The Generalized Video Traffic Model (GVTM) from Chapter 4 is adopted to model MPEG4 video streaming traffic. GVTM is extended with a rate control mechanism from [100] so that output bit rate can be configured.

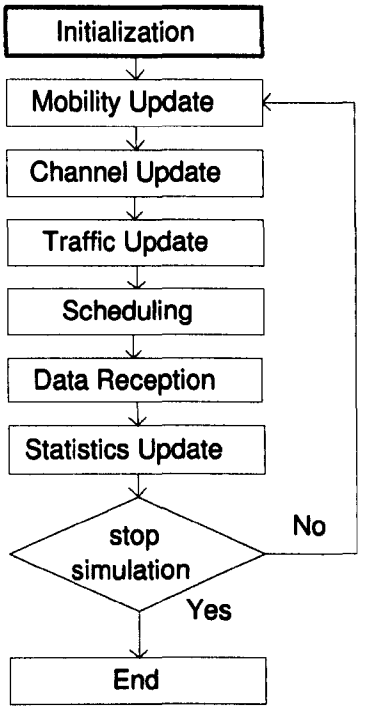


Figure 5.19: Simulation manager operation

5.4.6 Timer

A timer is used to keep track of the simulation time elapsed. This is useful for various calculations such as packet delay and average bit rate. The timer is incremented with a fixed value for every Time Transmission Interval (TTI). The increment step considered in the system level simulator is 2ms.

5.4.7 Simulation Manager

The simulation manager is the core of the system level simulator. It creates and manages all the components within the system. Fig. 5.19 shows the operation of the simulation manager. At the initialization stage, the simulation manager creates all the system components (e.g. mobile stations and base stations) and

initializes them with the default parameter values. During the simulation, the following steps are iterated:

- Mobility update: Update the location of every mobile station
- Channel update: Update the channel of every mobile station. The path loss and shadowing loss of mobile stations are calculated. Multipath channel gain is obtained from pre-calculated traces.
- Traffic update: The traffic model generates traffic into the base station queue.
- Scheduling: Base station selects a group of mobile stations for transmission based on their queue and channel conditions. Power, modulation and channel code rate are selected for individual mobile stations. Transmission of data block begins.
- Data reception: Calculate the SIR of the mobile station. Map SIR of QAM symbols to effective AWGN SIR using the EESM method. Block error rate is calculated given the effective AWGN SIR. Decide if the data block is successfully transmitted considering the block error rate.
- Loop back until simulation duration expires.

At the end of the simulation, all the statistics such as packet delay, and packet delay violation probability are collected and saved in an output file.

5.5 Conclusions

In this chapter, the design and implementation of a link level simulator and a system level simulator are presented. Using the link level simulator, the block error rate (BLER) performance curves are obtained for different Modulation and Coding Schemes (MCS). BLER performance curves are useful for the modelling of error performance in system level simulations. Link level to system level simulator interface called Exponential Effective SIR Mapping (EESM) is introduced. Finally, the development of a system level simulator is described. Both simulators will be used in succeeding chapters for the study of multimedia resource allocation and scheduling.

Chapter 6

Wireless Multimedia Resource Allocation and Scheduling

6.1 Introduction

Orthogonal Frequency Division Multiplexing (OFDM) is a strong candidate for 4G wireless air interface due to its inherent robustness against frequency selective channel fading. It is being considered within several standardization bodies such as 3GPP, IEEE 802.11, 802.16 and 802.20 standards. Qualcomm has also acquired Flarion Technologies [101] to strengthen their OFDM portfolios despite the fact that they are the pioneer of Code Division Multiple Access (CDMA) wireless technology. This shows the popularity of OFDM air interface and it is generally perceived that OFDM will be the next generation air interface for 4G systems. As OFDM systems are widely deployed, it has become urgent to study OFDM specific resource allocation and scheduling algorithms to maximize system resource utilization. Meanwhile, various developments within audiovisual technologies such as the MPEG4 and H.264 have enabled the transmission of video over wireless environments. In particular, video is more resilient to errors

due to channel fading as well as being able to adapt its source rate to match the time varying channel bandwidth. Wireless video streaming is now seen to be a revenue generating service by the cellular and broadband wireless operators. Efficient resource allocation algorithms are required to partition a pool of resources to a group of incoming users such that systemwide user satisfaction is obtained. In light of advances in both areas, the objective of this chapter is to develop efficient resource allocation and scheduling algorithms for transmission of video streaming traffic over OFDM based systems.

The chapter is divided into three parts. The first part introduces the system model. The second introduces a cross layer resource allocation called the Genetic Algorithm Based Resource Allocation (GABRA) technique. GABRA jointly considers user level performances and channel conditions to optimize multiple video streaming sessions so that systemwide user satisfaction is achieved. The final part of this chapter proposes a resource scheduling algorithm for delay sensitive video streaming traffic. This scheduling algorithm is called Minimum Queue Length Ratio Scheduler (MQLRS). GABRA and MQLRS are proposed for a Frequency Hopped Orthogonal Frequency Multiple Access (FH-OFDMA) system. FH-OFDMA [102] is considered due to its desirable properties such as elimination of intracell interference in CDMA and being capable of achieving frequency diversity and intercell interference diversity. Although FH-OFDMA does not exploit multiuser diversity gain, it allows frequency reuse of one and facilitates network deployment without time consuming frequency planning [93].

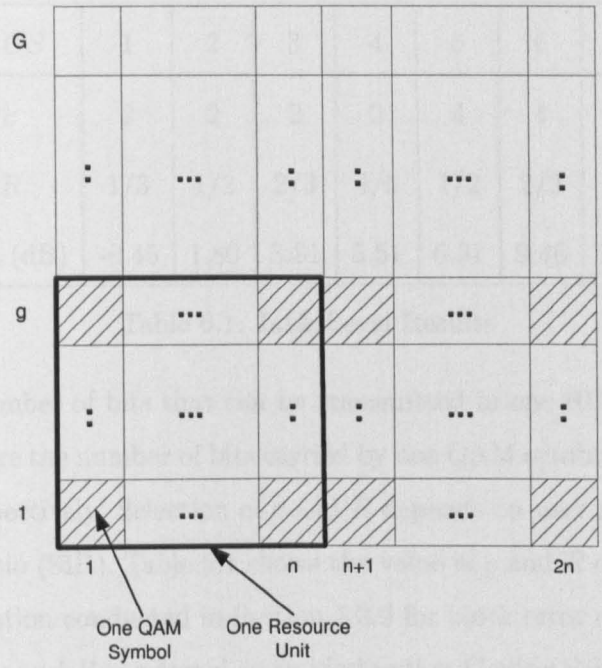


Figure 6.1: Time frequency resource in OFDM system

6.2 System Model

6.2.1 OFDMA Time Frequency Resource

Fig. 6.1 shows the time-frequency resources in an OFDMA system. In an OFDMA system, the frequency bandwidth is divided into G orthogonal subcarriers. The smallest unit of resource in the OFDMA system is one subcarrier carrying one Quadrature Amplitude Modulation (QAM) symbol. It is possible to group several subcarriers over several OFDM symbol periods as a single Resource Unit (RU). The basic RU considered in this chapter consists of g subcarriers spanning over n successive OFDM symbol periods as portrayed in Fig. 6.1. The Transmission Time Interval (TTI) is defined as the duration of one RU, T_{TTI}. Following on from previous descriptions, T_{TTI} can be calculated as nT_s where T_s is one OFDM symbol period. The total number of RUs within each TTI, Z, can be calculated

MCS	1	2	3	4	5	6	7
c	2	2	2	2	4	4	4
R	1/3	1/2	2/3	4/5	1/2	2/3	3/4
SIR (dB)	-0.45	1.80	3.81	5.51	6.91	9.46	10.81

Table 6.1: Link Level Results

as G/g. The number of bits that can be transmitted in one RU is $c \times R \times g \times n$ where c and R are the number of bits carried by one QAM symbol and the channel coding rate respectively. Selection of c and R depends on user received Signal to Interference Ratio (SIR). Table 6.1 shows the value of c and R obtained from the link level simulation conducted in Section 5.3.9 for block error rate of 0.01. Each combination of c and R is referred to as Modulation Coding Scheme (MCS). The total number of RUs within a time window S_{T_w} is

$$S_{T_w} = \frac{T_w}{T_{TTI}} \times \frac{G}{g}, \tag{6.1}$$

where T_s is one symbol period. Suppose that S_k is the number of RUs within T_w that is allocated to k^{th} user, the achievable bit rate B_k is

$$B_k = B(S_k) = \frac{S_k \times c_k \times R_k \times g \times n}{T_w}, \tag{6.2}$$

where subscript k indicates that c and R are user specific and depend on the received SIR of the individual user.

6.2.2 Frequency Hopping Using the Latin Square

The latin square method [92] is adopted for frequency hopping over G subcarriers. Eq. (6.3) illustrates an example of the latin square hopping matrix

$$\mathbf{H} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \\ 5 & 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 & 1 \\ 4 & 5 & 1 & 2 & 3 \end{bmatrix}. \tag{6.3}$$

Let $H_{i,j}$ be the element of the matrix in Eq. (6.3). Elements in the j^{th} column of the matrix, i.e. $H_{1,j} - H_{W,j}$, corresponds to the hopping pattern of subcarriers. W is the horizontal and vertical dimension of the matrix. The j^{th} column in $H_{i,j}$ corresponds to j^{th} OFDM symbol time. In the latin square matrix, the value of element is unique for any given row or column. For example, the value from 1 to 5 in Eq. (6.3) only appears once in any given row or column. In addition, when the hopping pattern reaches the last column of the matrix, it wraps around and restarts from the first column of the matrix. Hence the hopping patterns are periodic. When W is prime, there is a simple way of constructing a family of $W-1$ mutually orthogonal latin squares [92]. For $a = 1, \ldots, W-1$, the element of matrix H for a^{th} member in the family can be calculated from

$$H^a_{i,j} = (a \times i + j) \text{ modulo W}. \tag{6.4}$$

6.2.3 Overall Video Transmission System

The basic system model under study is depicted in Fig. 6.2. The Resource Allocation Unit (RAU) efficiently distribute RUs to users such that the total user satisfaction is maximized. The allocation decision is then passed to individual

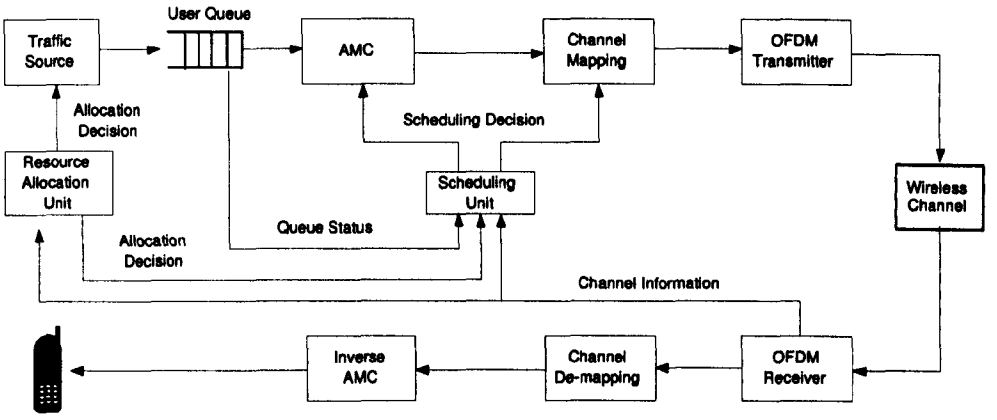


Figure 6.2: OFDMA system model

user applications as well as the scheduling unit. The Scheduling Unit (SU) dynamically selects users for transmission based on QoS parameters and queue conditions of all users. For each scheduled user, the SU selects transmission power, P_k, and MCS_k (i.e. c_k and R_k) based on user received SIR. Decisions of P_k, and MCS_k for each user are passed to the Adaptive Modulation and Coding (AMC) block where the user bit stream is turbo coded and transformed to QAM symbols. Finally, the channel mapping block maps QAM symbols onto RUs location using the latin square hopping pattern. The resulting RUs are sent over the wireless channel using the OFDM transmitter. In this thesis, equal power allocation is assumed for simplicity. MCS_k is adapted every T_w to slow channel variation which includes path loss and shadowing. A fading margin is added to absorb small scale fading.

6.3 Joint User- and Channel- Aware Resource Allocation

Error prone and time varying wireless channel characteristics pose challenging technical issues for resource allocation. In particular, it is highly inefficient to guarantee deterministic QoS to users as this requires excessive system resource to counteract the channel defect. For this reason, adaptive resource allocation is of interest in which the user application QoS is adapted to network or channel conditions [103]. For application such as MPEG4 video streaming, this does not create significant problems as MPEG4 video source rate can easily be adapted to wireless channel bit rate. MPEG4 source rate adaptation is achieved either by rate control or bitstream truncation (for FGS coded video) [104] although at the expense of video quality. At the same time, as the technology evolves, it has become clear that human factors or a user-centered approach assumed increasing importance for multimedia adaptation [63]. The author in [63] has claimed that adaptation should consider user-perceived quality to improve user satisfaction. In [65] [66], the authors have proposed user-awared single user video adaptation techniques. In [105], joint source and channel coding that considers user level performance has been introduced for the Adaptive Multi-Rate (AMR) voice codec. The author considers user level performances, i.e. mean opinion score (MOS), by using an ITU E-Model [105] for adaptation. The algorithms described previously are mainly for single user cases and omit time varying channel bandwidth behaviour. In this chapter, a resource allocation technique is proposed for multiple users with consideration for user level performances and time varying channel conditions. The algorithm is described in detail in the following sections.

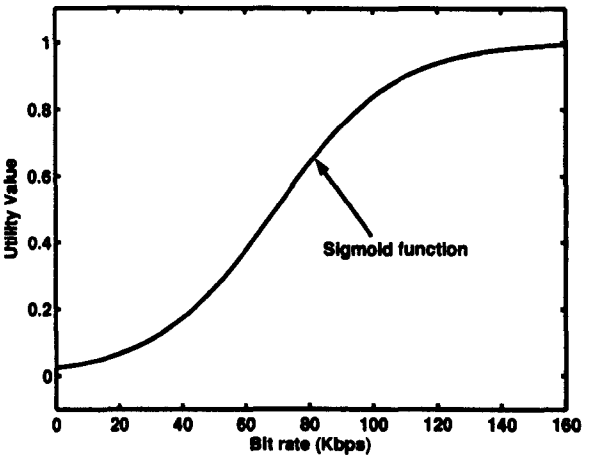


Figure 6.3: Utility functions for video services

6.3.1 Problem Formulation

Resource allocation problems are challenging in wireless environments due to unreliable and time varying channel conditions. As individual users experience different received SIR, the supportable MCS for each user is also different. Therefore, different users may require different numbers of RUs even for the same bit rate requirement. RAU seek to allocate system RUs to incoming users such that systemwide user satisfaction is achieved. A utility theory is used to measure user satisfaction for the resource allocation scheme. In utility theory, a mathematical function is used to map the allocated RUs to an utility value that quantifies the satisfaction of each user. For maximizing the satisfaction of all users, the problem can be formulated as

$$\max_{\mathbf{S}=\{S_1,\ldots,S_K\}} \quad \sum_k U(B(S_k)) \tag{6.5}$$

$$\text{subject to } \sum_k \; S_k \leq \alpha S_{T_w} \tag{6.6}$$

$$S_k \geq S_{k,low}, \tag{6.7}$$

where $\mathbf{S} = \{S_1, ..., S_K\}$ is the solution to Eq. (6.5) and $U(\cdot)$ is an objective function that maps the allocated bit rate $B_k = B(S_k)$ to a satisfaction value.

The first constraint indicates that the total allocated RUs should not exceed fraction α of total available RUs, S_{T_w}. The second constraint ensures user receive at least $S_{k,low}$ of slots. $S_{k,low}$ is calculated from the user specified minimum bit rate bound, $B_{k,low}$ using Eq. (6.2). The above formulation is general and can take many forms of objective functions. In video coding literature, the objective function often take the form of rate to distortion mapping [106]. In these works, optimization algorithms are used to minimize the total video distortion. The objective function may also be constructed by using a subjective evaluation such as the Mean Opinion Score (MOS). In this study, a sigmoid utility function [67] is assumed for rate adaptive MPEG4 video streaming. The sigmoid utility function can be represented as

$$U(B) = [1 + \exp(-5.36 \times 10^{-5} \times B + 3.7143)]^{-1}. \tag{6.8}$$

The sigmoid utility function is shown in Fig. 6.3. The sigmoid utility function is chosen as it correlates to human perceptual of quality well. The curve shape signifies the fact that the user perceived quality goes to zero when B_k falls below a certain value. In reverse, if B_k allocated is more than a certain value, the user perceived quality is not further improved. Sigmoid utility function is also discussed in [67] for video streaming. Although sigmoid utility function is assumed, a more sophisticated utility function can be used if it is available in the future. Utility function can be estimated by subjective evaluation during a test campaign. Estimation of utility function is beyond the scope of this thesis.

6.3.2 Genetic Algorithm Based Resource Allocation

A Genetic Algorithm Based Resource Allocation (GABRA) scheme is proposed to solve Eq. (6.5). The proposed GABRA algorithm is generalized and can be used for different shapes of utility function. The steps taken to solve Eq. (6.5) are

1. Let $\mathbf{\Omega}(m)$ be a set of solutions at the m^{th} generation. $\mathbf{\Omega}(m)$ has L solutions in the set and each solution can be represented as $\Omega_l(m) = \{\Omega_{l,m,1}, ..., \Omega_{l,m,K}\}$. In a genetic algorithm, a solution is also called chromosome while the number of solutions, L, is called the population size. First, initialize $\mathbf{\Omega}(0)$ with randomly generated values with $\Omega_{l,m,k}$ falling in the range $[S_{k,low}, \alpha S_{T_w}]$.

2. Calculate the fitness, $\sum_k U(B(\Omega_{l,m,k}))$, for all chromosomes within $\mathbf{\Omega}(m)$ using Eq. (6.8). It is possible that chromosomes within $\mathbf{\Omega}(m)$ violate the constraint in Eq. (6.6). In order to guide the genetic algorithm towards the feasible solution region, an approach called Niched-Penalty [107] is adopted to scale the fitness of chromosomes. The scaling is performed so that a chromosome that violates the constraint Eq. (6.6) will have a lower fitness when compared to chromosomes in feasible solution region. The scaling operation also ensures that a chromosome with less constraint violation has greater fitness value than a chromosome with higher constraint violation. The scaled fitness is calculated by

$$\mathbf{F}(\Omega_l(m)) = \begin{cases} \sum_k U(B(\Omega_{l,m,k})) + \max_l V_l & \text{if } V_l = 0 \\ -V_l + \max_l V_l & \text{if } V_l > 0 \end{cases}, \tag{6.9}$$

where V_l is the amount of constraint violation for l^{th} solution

$$V_l = \max(0, \sum_k \Omega_{l,m,k} - \alpha S_{T_w}). \tag{6.10}$$

3. Record the fittest solution.

4. Two parent chromosomes, Pc_1 and Pc_2, are selected for reproduction (or crossover) using the pairwise tournament selection method. During the first round, two chromosomes are selected using the roulette wheel method [108]. The chromosome with better fitness is kept as the first parent chromosome while the other is discarded. The same procedures are used to select the second parent chromosome in the second round.

5. Arithmetic crossover [108] is performed on two parent chromosomes, Pc_1 and Pc_2, to produce two children chromosomes, Chc_1 and Chc_2. The crossover operations are

$$\begin{aligned} Chc_1 &= \beta_1 \cdot Pc_1 + (1-\beta_1) \cdot Pc_2 \\ Chc_2 &= (1-\beta_1) \cdot Pc_1 + \beta_1 \cdot Pc_2 \end{aligned}, \tag{6.11}$$

where β_1 is a randomly generated value within $[0, 1]$.

6. Arithmetic mutation [108] is performed on each gene (each variable within the chromosome) of Chc_1 and Chc_2 with a probability of 0.01. The resulting chromosomes are denoted as Chc_1' and Chc_2'. Mutation is an important operation to avoid the search operation trapping at a local optima. Using gene k, $Chc_1(k)$, as an example, the mutation operation is

$$Chc_1'(k) = \beta_2 \cdot Chc_1(k) + (1-\beta_2) \cdot w_k, \tag{6.12}$$

where β_2 is a randomly generated value within $[0, 1]$. w_k is a randomly generated value bounded by $[S_{k,low}, \alpha S_{T_w}]$.

7. Repeat Step 4 to Step 6 until the total number of newly generated chromosomes reaches L. Newly generated chromosomes are then added to the existing population $\Omega(m)$. A total number of L chromosomes with low fitness are then removed to obtain the new population $\Omega(m+1)$.

8. Set $m = m + 1$, then repeat Step 2 to Step 7 until m reaches M. Return the best solution recorded in Step 3 and terminate the algorithm.

With the proposed GABRA method discussed above, the solution $\mathbf{S} = \{S_1, ..., S_K\}$, maximum systemwide user satisfaction can be calculated while at the same time satisfying individual user QoS requirement, $S_{k,low}$. Since user channel conditions vary over time, the number of RUs required by each user changes over time. RAU is executed frequently to ensure that maximum user satisfaction is always obtained.

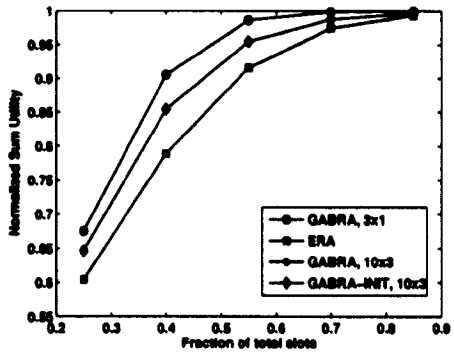
6.3.3 Performance of GABRA

Parameter	Setting
Cell Radius	1.6 km
Carrier Frequency	2 GHz
Base Station Power	43 dBm
Path Loss	$128.1 + 37.6 \times log_{10}(d)$
Shadowing Loss	Lognormal, 8dB variance
Multipath Channel	ITU Ped-A, 30km/h

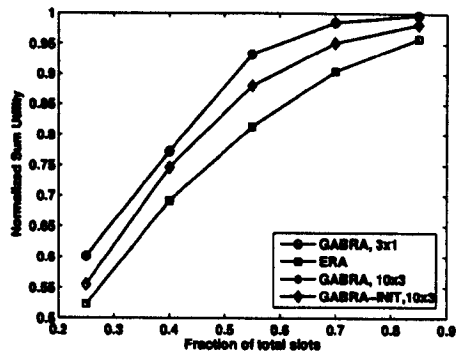
Table 6.2: System Level Parameters

The simulation parameters of the system model are described in this section. The parameters of OFDM transmitter and receiver are listed in Table 5.1. The system level simulation parameters are illustrated in Table 6.2. Path loss, shadowing loss and multipath channel models are described in detail in Section 5.4.3. Mobiles are placed uniformly within the cell. The $B_{k,low}$ that corresponds to $S_{k,low}$ is set to 16kbps for all video users. T_w is set to 100ms. Hence, the total RUs, S_{T_w}, is calculated to be 350 using Eq. (6.1). The simulation time is 30,000 TTI. Fig. 6.4 shows the plots of normalized sum utility averaged over simulation interval for GABRA. The Equal Resource Allocation (ERA) scheme is implemented for comparison purposes. In ERA, total RUs are equally distributed to all video users regardless of their channel conditions and user level performances. It can be seen that GABRA consistently outperforms ERA which is not user- and channel-aware. For 15 and 20 users, the number of slots required to achieve normalized sum utility of 0.9 for GABRA is about 15% less than ERA. For 25 and 30 users, the number of slots required to achieve normalized sum utility of 0.8 for GABRA is also about 15% less than ERA.

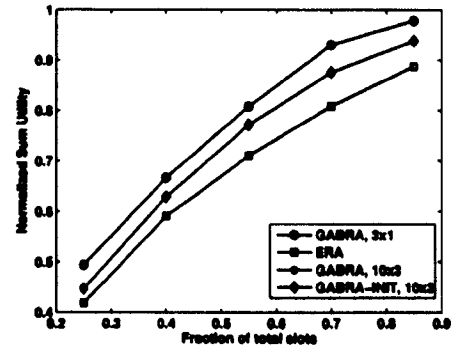
One of the requirement of GABRA is that it should perform optimally for differ-



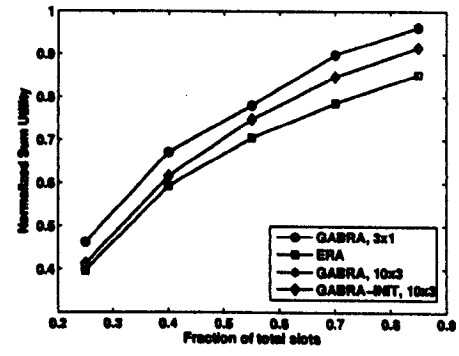
(a) 15 users



(b) 20 users



(c) 25 users



(d) 30 users

Figure 6.4: Time averaged system utility for different resource utilization

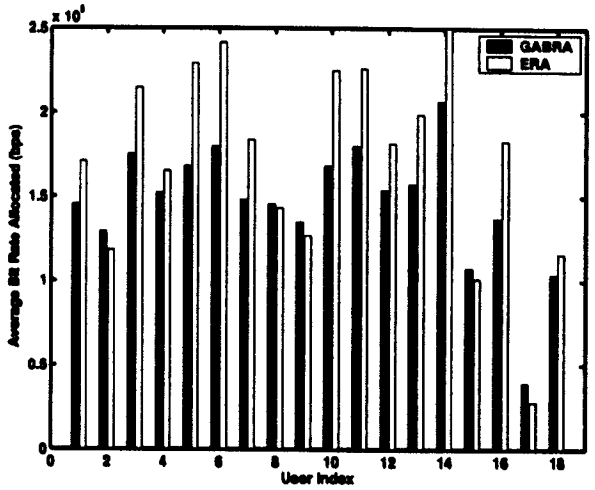


Figure 6.5: Time-averaged user allocated bit rate

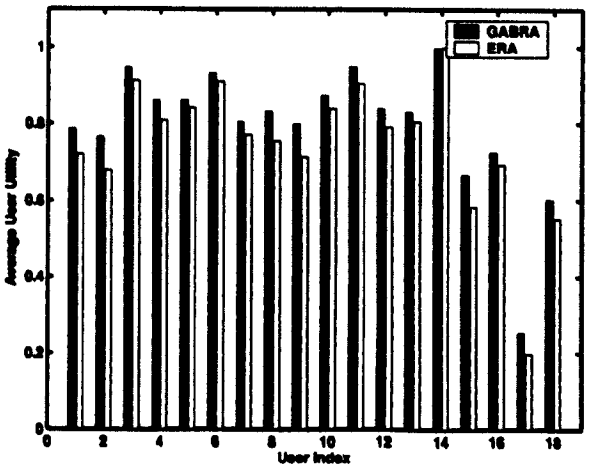


Figure 6.6: Time-averaged user utility

ent number of users, K. Thus, the number of generation, M, and population size, L, of genetic algorithm is not fixed and depends on the number of users. In order to examined this, values of $3K$, $5K$ and $10K$ are considered for M while values of K, $2K$ and $3K$ are considered for L. In Fig. 6.4, the legend "GABRA,10x3" represents $M = 10 \times K$ and $L = 3 \times K$. The figure shows two extreme cases, with $M = 10K$, $L = 3K$ and $M = 3K$ and $L = K$. The value of M and L can be as low as $3K$ and K without noticeable performance penalty. From the simulation, it can be concluded that GABRA performs well for different numbesr of user.

In Step 1 of Section 6.3.2, the solution pool is initialized randomly. There are potentially a high number of infeasible solutions and the solution may be far from a feasible solution region. Due to this, the solution may not converge fast enough to feasible solution region to achieve high quality solution. In order to examine this, Step 1 of Section 6.3.2 is extended to include an initializer to ensure that all the initial solutions in the pool does not violate the constraint in Eq. 6.6. For each randomly initialized solution, a gene is randomly selected and subtrated from a small fixed value. This procedure is repeated until the solution conform to Eq. 6.6. The result are shown in the figures as "GABRA-INIT, 10x3". It can be seen that initializing the solutions to feasible solution region is unnecessary and the gain is lower. This is possibly due to the fact that forcing the initial solutions to feasible region limits the diversity of the solutions and hence reduces the possibility of obtaining a high quality solution.

GABRA outperforms ERA because it is user- and channel- aware. By being aware of user level performance, GABRA can tradeoff RUs allocated to different users according to their channel conditions in order to achieve maximum systemwide user satisfaction. For example, a user may be in favourable channel conditions (i.e. high MCS) or operating near the saturation level of the utility function. In such

a situation, further allocation of RUs does not increase the satisfaction of that user. GABRA distributes these excessive RUs to compensate other users that are in poor channel conditions (i.e. low MCS) or low utility region to further increase the total system utility. This behaviour can be further observed in Fig. 6.5 and Fig. 6.6. Fig. 6.5 and Fig. 6.6 show the time averaged allocated rate and achieved utility respectively. In Fig. 6.5, some users experience lower throughput than ERA. This is because these users are in good channel conditions and supportable bit rate is high. However, allocating bit rate more than about 128kbps only brings marginal gain to the system utility. GABRA intelligently allocate these excessive resource from users with good channel conditions to users operating at lower utility. Hence, the user shows a higher time-averaged utility in Fig. 6.6 although the actual allocated throughput may be lower.

6.4 Resource Scheduling for Delay Sensitive Multimedia Traffic

Early studies of resource scheduling in OFDMA systems have mainly concentrated on physical layer aspects such as throughput and total transmission power. One of the earliest work in this area appears to have been carried out by Wong *et al.* in [44]. Wong *et al.* formulated the resource scheduling problem as a total transmit power minimization problem. The same problem has been examined in [91] and some reduced complexity algorithms have been introduced. These algorithms achieved less transmission power or high throughput by adaptively assigning a subcarrier to a user with high channel gain. However, these algorithms have not considered the user QoS requirements.

More recently, there are attempts to introduce Generalized Processor Sharing

(GPS) based techniques to schedule OFDMA resources to guarantee user QoS [109] [110] [111]. GPS is technique that assign a positive weight, η_k, to user k such that the guaranteed bandwidth, B_k, is

$$B_k = \frac{\eta_k}{\sum_{i=1}^{I} \eta_i} \times \Psi, \tag{6.13}$$

where I is the total number of user in the system and Ψ is the total link transmission rate. Diao *et al.* [109] combined a GPS based packet scheduler with channel aware algorithms to improve total system throughput and packet error rate. Zhang *et al.* [110] have considered a cross layer technique to jointly optimize bandwidth and power allocation while considering fairness guarantee. Cai *et al.* [111] considered a throughput maximization resource algorithm with fairness guarantee. Although GPS based schedulers described above achieved QoS guarantee, they assumed deterministic channel capacity, Ψ. In contrast, new generation wireless air interfaces feature multiple link bit rate characteristics due to adaptive modulation. This violates the basic assumption of GPS and hence the GPS based algorithms may not be suitable for a system with adaptive modulation. In some cases, GPS based schedulers feature high complexity due to intensive computation of virtual time for each packet and emulation of error free flows [110] [58].

A packet scheduler called Minimum Queue Length Ratio Scheduler (MQLRS) is proposed in the following sections for the transmission of delay sensitive traffic over OFDMA system. MQLRS aims to minimize delay experienced by individual according to the requested QoS. While GABRA allocate a number slots to individual user, MQLRS aim to distribute these slots to user dynamically according to the delay requirement of individual users. MQLRS can be used in new generation wireless air interface with adaptive modulation as there is no assumption of deterministic capacity as in GPS based schedulers. In summary, the scheduling problem is formulated as an optimization problem where the aim is to enforce

throughput guarantee within an individual users declared delay window. In addition, MQLRS is able to distribute the delay violation probability to all the users fairly.

6.4.1 Problem Formulation

In this section, downlink resource scheduling for multiple users with delay-sensitive multimedia traffic in OFDMA system is considered. The target of the proposed MQLRS aims at satisfying user bandwidth and delay requirements. The formulation of MQLRS is described as follows. Suppose that the user required queue length violation probability be

$$\Pr(Q_k > Q_{k,\max}) \leq \delta_k, \tag{6.14}$$

where Q_k and $Q_{k,\max}$ are respectively the instantaneous queue length and the target queue length of user k. $Q_{k,\max}$ can be calculated using

$$Q_{k,\max} = B_k \times D_{k,\max}, \tag{6.15}$$

where B_k and $D_{k,max}$ are the user bit rate and delay requirement. During heavy traffic load conditions, the following approximation holds [112]

$$\Pr(Q_k > Q_{k,\max}) \approx \exp(-\frac{Q_{k,\max}}{\tilde{Q}_k}), \tag{6.16}$$

where $\tilde{Q}_k$ is the mean queue length of user k. Equating Eq. (6.16) to δ_k, $\tilde{Q}_k$ can be estimated as

$$\tilde{Q}_k = \frac{-Q_{k,\max}}{\ln(\delta_k)}. \tag{6.17}$$

Thus user k will have queue length violation probability less than δ_k if its Q_k is always less than $\tilde{Q}_k$. Based on this rationale, the scheduling problem is formulated as sum queue length ratio minimization problem

$$\min \sum_k \frac{Q_k}{\tilde{Q}_k}. \tag{6.18}$$

The above formulation aims to ensure that the queue length of all users is no more than their respective target mean queue length, $\tilde{Q}_k$. Thus the proposed scheduler is called Minimum Queue Length Ratio Scheduler (MQLRS). MQLRS is designed in conjunction with a credit based technique. In the credit based technique, a user may transmit only when the user has accumulated sufficient credit. This provides flow isolation between users and guarantee minimum throughput if the scheduler is properly designed. In MQLRS, a counter V_k starts to accumulate credits when the queue is not empty. In every TTI, V_k is incremented with the number bits a user is entitled to send i.e. $B_k \times T_{TTI}$. When user k is served, V_k is decremented by the amount of bits transmitted. V_k is reset to zero when the queue is empty. Since V_k build up in the same way as user traffic is queued, Eq. (6.18) can be rewritten as

$$\min \sum_k \frac{V_k}{\tilde{V}_k}, \tag{6.19}$$

where V_k is instantaneous credit value while $\tilde{V}_k$ is the target average credit value calculated using Eq. (6.17). It can be seen that MQLRS is simple and only requires a counter, V_k, to keep track of the service accumulation.

6.4.2 Minimum Queue Length Ratio Scheduler (MQLRS)

The scheduling problem is formulated as a queue length ratio minimization problem in the previous section. It is not difficult to see that priority can be given to users with greatest $V_k/\tilde{V}_k$ for transmission in order to solve Eq. (6.19). Let Ψ and Ω be the set of users with a non-empty queue and the set of users to be scheduled respectively. Ω is obtained as follows

Set $\Omega \leftarrow \varnothing$

Set $\Psi \leftarrow \{k : Q_k > 0,\ \forall k \in \{1, ..., K\}\}$

while $\#\Omega < Z$ **do**

$\gamma^* \leftarrow \arg\max_{\gamma \in \Psi} \frac{V_\gamma}{\tilde{V}_\gamma}$

$\Omega \leftarrow \Omega \cup \{\gamma^*\}$

$\Psi \leftarrow \Psi \backslash \{\gamma^*\}$

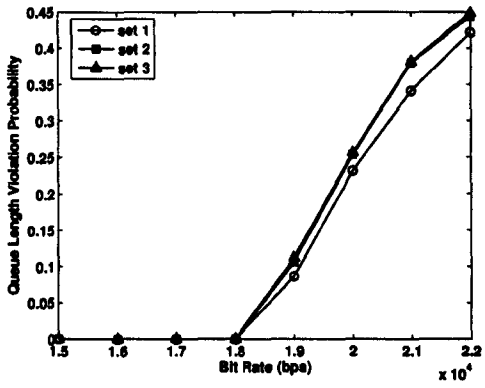
end while

where $\#\Omega$ is the cardinality (the number of elements) of the set Ω and Z is defined earlier in Section 6.2.1. At each step, user k with maximum $V_k/\tilde{V}_k$ is selected from the set of users with non-empty queue, Ψ. The selected user is then added to set, Ω, which will be scheduled at a later stage. The selection stops when the number of users selected reaches Z. From the above steps, one can observe that MQLRS tries to enforce throughput guarantee B_k over individual delay time window $D_{k,max}$ to keep user experienced delay low. In order to illustrate this idea clearly, suppose there are only two users, user 1 and user 2 in the system. User 1 and user 2 have the same bit rate requirement, B_1 and B_2, but with different delay requirements, $D_{1,max}$ and $D_{2,max}$ where $D_{1,max} < D_{2,max}$. Let the target average credit value of user 1 and user 2 be $\tilde{V}_1$ and $\tilde{V}_2$ respectively. It is known that $\tilde{V}_1 < \tilde{V}_2$ since $D_{1,max} < D_{2,max}$. Further assume that V_1 is equal to V_2 at some time instant. It can be seen that user 1 will always be given priority for transmission when compared to user 2 since $V_1/\tilde{V}_1 > V_2/\tilde{V}_2$. This shows how MQLRS gives transmission priority to the user with an urgent deadline over user with less urgent deadline.

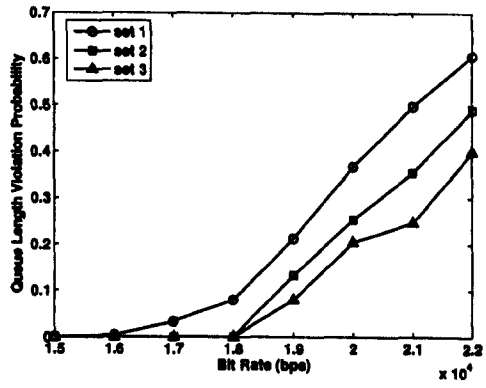
Let δ_k in Eq. (6.17) be set to an equal value for all the users, then MQLRS is also able to distribute the credit queue length violation probability to different users equally. This is because MQLRS maintains the same $V_k/\tilde{V}_k$ for all the users during congestion. This in turn means that the experienced average credit queue length, V_k^*, relative to $V_{k,max}$ for all users is similar. When $\tilde{V}_k$ is substituted by V_k^* in Eq. (6.16), it can be seen that the estimated credit queue length violation probability will be close for all the users.

6.4.3 Performance of MQLRS

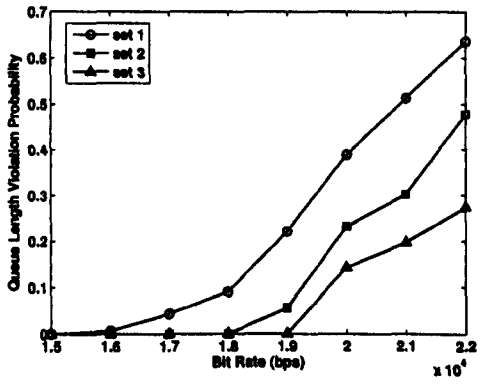
The simulation setting for evaluating MQLRS is the same as the simulation setting in Section 6.3.3. Video traffic generator from Section 5.4.5 is implemented to simulate video users. The performance of MQLRS is measured in terms of credit queue length violation probability and average packet queuing delay. In credit queue length violation probability, the probability of current credit queue length, V_k, greater than target queue length, $V_{k,\max}$, is measured. $V_{k,\max}$ of Eq. (6.15) defines the maximum acceptable delay window for a given user. Two other algorithms called Maximum Credit First Scheduler (MCFS) [113] and Maximum Delay First Scheduler (MDFS) [114] are implemented for comparisons. In MCFS, users with the longest credit queue are given priority for transmission. As for MDFS, users with the longest delay (which is credit queue divided by target bit rate) are scheduled for transmission first. Note that MCFS and MDFS were originally proposed for MultiCode-CDMA systems, but the principle of scheduling still remains the same during implementation. The same principle of scheduling in this context means that the scheduling cost function or criteria for selecting a user for transmission, e.g. maximum queue length first or maximum delay first, are still the same. Simulations are run over 30,000 TTIs. There are 3 sets of video users in the system with 10 video users in a set. The first and second sets of video users have bit rate of 100kbps and 150kbps respectively. Delay requirements of video user set 1 and set 2 are 50ms and 225ms respectively. The third set of video users has a bit rate range of 150kbps to 220kbps with increment steps of 10kbps. The delay requirements of video user set 3 is 500ms. Fig. 6.7 shows the average credit queue length violation probability for video user sets 1, 2 and 3 for MQLRS, MDFS and MCFS. It can be seen in Fig. 6.7(a) that the average credit queue length violation probabilities of MQLRS for video user sets 1, 2 and 3 closely follow each other. This verifies the theory of queue length violation probability sharing described in Section 6.4.2. As for MDFS and MCFS, the average credit



(a) MQLRS

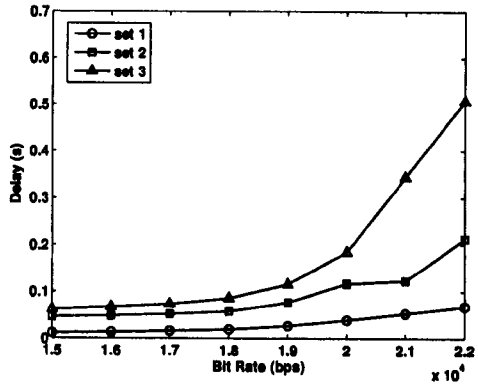


(b) MDFS

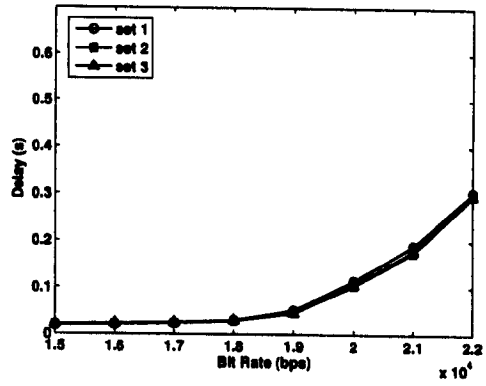


(c) MCFS

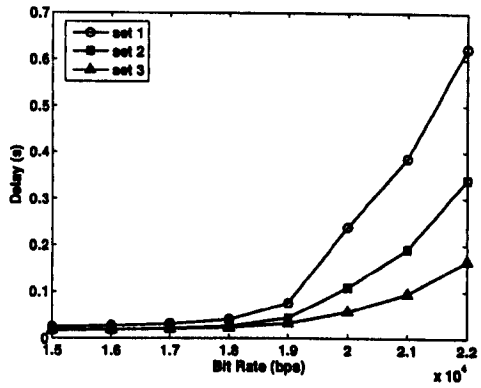
Figure 6.7: Credit queue length violation probability



(a) MQLRS



(b) MDFS



(c) MCFS

Figure 6.8: Packet queuing delay

queue length violation probability is unevenly distributed, with the users having shorter delay requirements experiencing higher queue length violation probability. Fig. 6.8 shows the average packet queuing delay of all users. The delay of video user sets 1, 2, and 3 for MQLRS increase according to their delay requirements as expected. As for MDFS, it distributes the queuing delay to all video users equally and omits the delay requirement of users. Finally, MCFS unfairly schedules the user with the longest credit queue length for transmission first. This scheme favours the user with high bit rate and thus less delay is experienced by video users in set 3 when compared to users in set 1. From these observations, it can be concluded that MQLRS is able to minimize delay according to the user declared delay window and fairly distributes queue length violation probability to all users.

6.5 Conclusions

In this chapter, resource allocation and scheduling for video streaming in OFDMA system are discussed. First, the basic resource structure in OFDMA system is established. Using a utility theory, the resource allocation problem is formulated as an optimization problem where the objective is to maximize systemwide user perceived quality. An efficient Genetic Algorithm Based Resource Allocation (GABRA) is then proposed to solve the problem. Simulation results show that GABRA outperforms the approach that disregards the channel conditions and user level performance. A resource scheduling algorithm called Minimum Queue Length Ratio Scheduler (MQLRS) is then proposed for the transmission of delay sensitive traffic over OFDMA system. MQLRS is designed in conjunction with a credit based technique where a credit counter is used to accumulate service credit. This is to provide flow isolation between users. Using the delay declared by individual users, a maximum credit accumulation limit can be calculated.

MQLRS gives priority to users with credit accumulation approaching their maximum credit accumulation limit for transmission. In other words, MQLRS tries to enforce throughput guarantee to users within their declared delay window. In addition, it has been shown that MQLRS is able to fairly distribute queue length violation probability among the transmitting users. Simulation results confirm that MQLRS is capable of minimizing delay and delay violation probability.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The achievements of the project during the period of study are described in the following paragraphs.

Chapter 3: Chapter 3 studies the statistical characteristics of MPEG4 encoded video traffic in detail. The basic properties of video traffic are marginal distribution of frame size, autocorrelation of frame size and cross correlation between different frame types. The marginal distribution can be modelled using a hybrid Gamma/Pareto probability distribution. Then a MultinoMial method (MM) is proposed to model the cross correlation between different frame types. Based on the MM, two video traffic models are proposed. The first model combines an autocorrelation modelling technique called Spatial Renewal Process (SRP) with MM (henceforth named SRP-MM). The second model combines an autocorrelation modelling technique called Nested AutoRegressive (Nested-AR) with MM (henceforth named Nested-AR-MM). The proposed SRP-MM and Nested-AR-MM are shown to capture the marginal distribution and autocorrelation of

empirical frame size accurately. Simulation results have shown that the proposed SRP-MM and Nested-AR-MM predict the empirical queuing performance with high accuracy and outperform some existing models.

Chapter 4: Chapter 4 presents a Generalized Video Traffic Model (GVTM) as an extension to the work in Chapter 3. Using an adaptable frame size model, GVTM is capable of generating video traffic for different quantization parameter set in real time. Thus, GVTM is more flexible and avoids the time consuming model parameters re-estimation process. Simulation results have shown that the GVTM accurately captures the inherent characteristics of video traffic i.e. frame size marginal distribution, autocorrelation of frame size and cross correlation between frame types. GVTM is shown to predict the queuing performances of empirical traffic with high accuracy for different buffer sizes and bandwidth utilizations for wide range of quantization parameter sets.

Chapters 5 and 6: In Chapter 5, an OFDM link level simulator is built to simulate the BLock Error Rate (BLER) curves for different Modulation and Coding Schemes (MCS). The Chapter then describes the implementation of a system level simulator. The system level simulator utilizes the BLER curves to model the error performances of OFDM system. An Exponential Effective SIR Mapping (EESM) technique is used for interfacing the link level simulator to the system level simulator. Chapter 6 studies the topic of multimedia resource allocation and scheduling in the Frequency Hopped Orthogonal Division Multiple Access (FH-OFDMA) system using the system level simulator. The first part of the chapter studies the resource allocation problem. Using a utility theory, the resource allocation problem is formulated as an optimization problem where the objective is to maximize systemwide user perceived quality. An efficient Genetic Algorithm Based Resource Allocation (GABRA) that jointly considers the channel condition and

user level quality is proposed. GABRA trades resources between users considering their channel condition in order to achieve systemwide user perceived quality. Simulation results have shown that GABRA outperforms an Equal Resource Allocation (ERA) approach that disregards the channel condition and user level quality. The second part of Chapter 6 introduces a scheduling algorithm called Minimum Queue Length Ratio Scheduler (MQLRS) for the transmission of delay sensitive multimedia traffic. It achieves low delay transmission by enforcing the throughput guarantee within the user declared delay window. MQLRS is also shown to fairly distribute queue length violation probability to all transmitting users. Simulation results have confirmed that MQLRS is capable of minimizing delay and delay violation probability.

7.2 Future Work

This section describes some of the issues, in the author's view, that remain to be tackled in the future.

Chapters 3 and 4 have examined the topic of MPEG4 video traffic modelling. However, as audiovisual technology advances, new single layer(e.g. H.264) and multi-layer video codecs (e.g. FGS) are being introduced. Hence, new video traffic models are required for these codecs for the evaluation of networking performances. Also, it is interesting to investigate the impact of channel coding such as turbo coding or multiple description coding on the video traffic. Hence, video traffic modelling is an ongoing research topic. The more urgent research topic is the modelling of multiple description coded video traffic as it can be used for the study of multiple path diversity. Multiple description coding [115] [116] encodes the data into several layers and each layer is independently transmitted over different transmission path to achieve diversity.

Multimedia resource allocation and scheduling problems in wireless environments are examined in Chapters 6 and 7. The topic that requires attention is how to define an accurate perceptually motivated quality metrics as opposed to conventional objective quality metric e.g. Peak Signal to Noise Ratio (PSNR) for video coding. A properly defined perceptually motivated quality metric enables one to study the effect of source parameters and network parameters on the user observed quality. This allows various resource allocation algorithms to be designed based on user perceived quality (e.g. GABRA algorithm in Chapter 7).

Multimedia resource allocation and scheduling problems in combined multiple antennas and orthogonal division frequency multiple access (MIMO/OFDMA) systems are still an open problem. This is due to the high degree of freedom in MIMO/OFDMA system parameters and finding an optimal algorithm is difficult. Thus, vast amounts of research efforts are required to find sub-optimal or reduced complexity algorithms. An example question to ask is: what frequency band, power, modulation and coding rate should be assigned to users in order to satisfy their QoS requirements ? These decisions are to be made considering the physical location of users, channel condition and diverse users' QoS requirements. The physical locations of users are important in resource MIMO/OFDMA allocation algorithms to avoid allocating users close to each other on the same frequency band and time slot that cause unnecessary power interference. It should be noted that careful design of resource allocation and scheduling algorithms are critical as certain MIMO schemes (e.g. transmit diversity) may have negative impact on the capacity gain of opportunistic scheduling [51].

It is noted that the resource allocation and scheduling algorithms are air interface specific. However, it is currently not clear which interface, e.g. OFDM-TDMA,

OFDMA, FH-OFDMA etc, should be adopted for next generation wireless systems. In this sense, more work is required to evaluate different interfaces in terms of their capacity performance, robustness against multicell interference and simplicity of deployment [117] [118]. Given the chosen air interface, resource allocation and scheduling algorithms can be further optimized under the scheme.

MIMO spatial multiplexing schemes provide multiple uncorrelated spatial channels for transmissions. As a result, path diversities can be exploited to achieve transmission robustness. However, additional dimension of antenna and time varying bit rate on each antenna represent additional degree of freedoms for optimization. It is interesting to investigate the transmission of video with scalable coding over multiple uncorrelated spatial channels. Scalable coding is considered because video source rate can be varied easily to match the channel bandwidth. In this case, possible options available for packetization of video data. For example, one may alternatively transmit each bit over different antenna or individual video packet over different antenna in round robin fashion or assign video packet with visual importance over more reliable antenna and so forth. Power can be unequally distributed among antennas according to reliability required for each video stream. It is interesting to investigate the performances of various schemes utilizing channel diversity on the video quality of scalable codec. Also, the schemes proposed should adapt easily to consider different number of antennas since adaptive allocation of antennas to user is possible in multiple antenna systems. Alternatively, Multiple Description Coding (MDC) [115] can be used to utilized the channel diversity provided by multiple antennas system. MDC is a class of techniques that is able to encode video sequence into multiple equally important descriptions. Congestion or fading on individual channel does not disrupt the video service completely since decoding process is still possible given that other descriptions arrived correctly. Although uncorrelated spatial channels

are assumed, but there is possibility that the spatial channels are not independent. The schemes proposed above should be evaluated for its robustness against channel correlation and improvements are to be incorporated wherever possible.

Appendix A

Author's Publications

1. C. H. Liew, C. K. Kodikara, A. M. Kondoz, "Modelling of MPEG-4 Encoded VBR Video Traffic", IEE Electronics Letters, vol 40, March 2004, pp. 355-357

2. C. H. Liew, C. K. Kodikara, A. M. Kondoz, "A Generalized Video Traffic Model for MPEG Encoded Model", IEEE 62nd Vehicular Technology Conference, US Texas Dallas, September 2005, pp. 1854-1858

3. C. H. Liew, C. K. Kodikara, A. M. Kondoz, "MPEG Encoded Variable Bit Rate Video Traffic Modelling", IEE Proceedings of Communications, vol 152, October 2005, pp. 749-756

4. C. H. Liew, C. K. Kodikara, A. M. Kondoz, "Resource Allocation Framework for Video Streaming Over Wireless OFDMA Systems", Joint NEWCOM-ACoRN Workshop Vienna, Sept 2006, submitted

5. C. H. Liew, A. M. Kondoz, M. Barbera, G. Schembra, "An Analytical Model for MPEG4 Transmission in Wireless Networks", Joint NEWCOM-ACoRN Workshop Vienna, Sept 2006, submitted

Appendix B

Mathematical Definition

B.1 Probability Distributions

B.1.1 Gamma Distribution

The Gamma distribution is:

$$F_{\Gamma}(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^x x^{\alpha-1} e^{\frac{-x}{\beta}} dx, \tag{B.1}$$

where $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$. The parameters β and α can be estimated as

$$\hat{\beta} = \frac{\hat{\sigma}^2}{\hat{\mu}} \tag{B.2}$$

$$\hat{\alpha} = \frac{\hat{\mu}}{\hat{\beta}}, \tag{B.3}$$

where $\hat{\sigma}^2$ and $\hat{\mu}$ are the estimated variance and the estimated mean of the frame activity. A numerical method [119] is adopted for the generation of Gamma random variables since it is difficult to obtain a close-formed equation for the inverse of (B.1).

B.1.2 Pareto Distribution

The Pareto distribution is:

$$F_P(x) = 1 - \left(\frac{k}{x}\right)^{\lambda}, \tag{B.4}$$

where k and λ are called the mode and the shape of pareto distribution.

B.2 Probability Integral Transform

Probability Integral Transform [120] says that if X is a random variable with continuous distribution function $F_X(x)$, then $U = F_X(X)$ is uniformly distributed over the range of (0,1). The proof is as follows:

$$\begin{aligned} P[U \leq u] &= P[F_X(X) \leq u] \\ &= P[X \leq F_X^{-1}(u)] \\ &= F_X(F_X^{-1}(u)) \\ &= u, \end{aligned} \tag{B.5}$$

where u is in the range of (0,1).

References

[1] (2005) Third generation partnership project. [Online]. Available: http://www.3gpp.org/

[2] (2005) Mobile worldwide interoperability for microwave access (WIMAX). [Online]. Available: http://www.ieee802.org/16/tge/

[3] L. Z. Ribeiro and L. A. DaSilva, "A framework for the dimensioning of broadband mobile networks supporting wireless internet services," *IEEE Personal Commun. Mag.*, vol. 9, pp. 6–13, June 2002.

[4] J. W. Roberts, "Traffic theory and the internet," *IEEE Commun. Mag.*, vol. 39, pp. 94–99, Jan 2001.

[5] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks," *IEEE J. Select. Areas Commun.*, vol. 23, pp. 1056–1066, May 2005.

[6] H. Ahn, J.Kim, S. Ching, B. Kim, and B. Choi, "A video traffic model based on the shifting-level process," in *Proc. of Infocomm*, 2000, pp. 1036–1045.

[7] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1142–1153, May 2005.

[8] *Packet Switched Conversational Multimedia Applications; Default Codecs*, 3GPP Std., Rev. 6.0.0, June 2003.

[9] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *Proc. of Conference on Communications Architechtures, Protocols and Applications*, London, UK, 1994, pp. 269–280.

[10] O. Cappe, E. Moulines, J. C. Pesquet, A. Petropulu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Processing Mag.*, vol. 19, pp. 14–27, May 2002.

[11] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level," *IEEE/ACM Trans. Networking*, vol. 5, pp. 71–86, Feb 1997.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic," in *Proc. of Conference on Communications Architectures, Protocols and Applications*, Sep 1993, pp. 183–193.

[13] W. Willinger, V. Paxson, R. H. Riedi, and M. S. Taqqu, *Theory and Applications of Long-Range Dependence.* Birkhauser, 2002.

[14] M. Grossglauser and J.-C. Bolot, "On the relevance of long-range dependence in network traffic," *IEEE/ACM Trans. Networking*, vol. 7, pp. 629–640, Oct 1999.

[15] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent packet traffic," *IEEE/ACM Trans. Networking*, vol. 4, pp. 209–223, April 1996.

[16] W. Stallings, *High-Speed Networks and Internets: Performance and Quality of Service*, 2nd ed. Prentice Hall, 2002, pp. 186–187.

[17] X. Huang, Y. Zhou, and R. Zhang, "A multiscale model for MPEG-4 varied bit rate video traffic," *IEEE Trans. Broadcast.*, vol. 50, pp. 323–334, Sept 2004.

[18] R. Riedi and J. Levy-Vehel, "TCP traffic is multifractal: A numerical study," *Technical Report RR-3129, INRIA, France*, 1997.

[19] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "The multiscale nature of network traffic: discovery, analysis and modelling," *IEEE Trans. Signal Processing*, vol. 19, pp. 28–46, May 2002.

[20] A. Nogueira, P. Salvador, R. Valadas, and A. Pacheco, "Modeling network traffic with multifractal behavior," *Telecommunication Systems Journal*, pp. 339–362, October 2003.

[21] A. Adas, "Traffic models in broadband networks," *IEEE Commun. Mag.*, vol. 35, pp. 82–89, July 1997.

[22] A. Rueda and W. Kinsner, "A survey of traffic characterization techniques in telecommunication networks," in *Proc. IEEE Canadian Conference on Electrical and Computer Engineering*, Calgary, Canada, May 1996, pp. 830–833.

[23] D. Jagerman, B. Melamed, and W. Willinger, *Stochastic Modeling of Traffic Processes.* CRC Press, 1998, pp. 271–320.

[24] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834–844, July 1988.

[25] P. R. Jelenkovic, A. A. Lazar, and N. Semret, "The effect of multiple time scales and subexponentiality in MPEG video streams on queuing behavior," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1052–1071, August 1997.

[26] J. Roberts, U. Mocci, and J. Virtamo, *Broadband Network Teletraffic: Final Report of Action COST 242.* Springer, 1996, pp. 7–46.

[27] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control.* Holden Day, 1970, pp. 46–125.

[28] M. Krunz and S. K. Tripathi, "On the characterization of VBR MPEG streams," in *Proc. of ACM SIGMETRICS Conference*, vol. 1, Washington, USA, May 1997, pp. 192–202.

[29] A. Golaup and A. H. Aghvami, "Modelling of MPEG4 traffic at GOP level using autoregressive processes," in *Proc. of IEEE VTC*, vol. 2, Vancouver, Canada, September 2002, pp. 854–858.

[30] J. Liu, Y. Shu, L. Zhang, F. Xue, and O. Yang, "Traffic modeling based on FARIMA models," in *Proc. IEEE Canadian Conference on Electrical and Computer Engineering*, Alberta, Canada, May 1999, pp. 162–167.

[31] D. Liu, E. I. Sara, and W. Sun, "Nested auto-regressive processes for MPEG-encoded video traffic modeling," *IEEE Trans. Circuits Syst.*, vol. 11, pp. 169–183, February 2001.

[32] S. Ma and C. Ji, "Modeling heterogeneous network traffic in wavelet domain," *IEEE/ACM Trans. Networking*, vol. 9, pp. 634–649, October 2001.

[33] ——, "Modeling video traffic in the wavelet domain," in *Proc. of INFOCOM*, vol. 1, San Francisco, CA USA, March 1998, pp. 201–208.

[34] ——, "Modeling video traffic using wavelets," *IEEE Commun. Lett.*, vol. 2, pp. 100–103, April 1998.

[35] O. Lazaro, D. Girma, and J. Dunlop, "A wavelet-based video traffic model for real-time generation of self-similar traffic," in *4th European Personal Communications Conference (EPMCC)*, Vienna, Austria, Feb 2001.

[36] N. Ansari, H. Liu, Q. Shi, and H. Zhao, "On modeling MPEG video traffics," *IEEE Trans. Broadcast.*, vol. 48, pp. 337–347, December 2002.

[37] M. Krunz, *"Statistical Multiplexing" in Encyclopedia of Electrical and Electronic Engineering.* John Wiley and Sons, 1999, pp. 479–492.

[38] H. Hlavacs, G. Kotsis, and C. Steinkellner, "Traffic source modelling," *Technical Report No. TR-99101, Institute of Applied Computer Science and Information Systems, University of Vienna, Austria*, 1999.

[39] M. Krunz and A. M. Makowski, "Modeling video traffic using M/G/inf input processes: A compromise between markovian and LRD models," *IEEE J. Select. Areas Commun.*, vol. 5, pp. 733–748, June 1998.

[40] M. Peng and W. Wang, "A framework for investigating radio resource management algorithms in TD-SCDMA systems," *IEEE Commun. Mag.*, vol. 43, pp. S12–S18, June 2005.

[41] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications - a key to gigabit wireless," *Proc. IEEE*, vol. 92, pp. 198–218, Febuary 2004.

[42] J. Tang and X. Zhang, "Cross-layer design of dynamic resource allocation with diverse QoS guarantees for MIMO-OFDM wireless networks," in *Proc. of the IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*, Taormina, Italy, June 2005, pp. 205–212.

[43] S. Catreux, V. Erceg, D. Gesbert, and R. W. Heath, "Adaptive modulation and MIMO coding for broadband wireless data networks," *IEEE Commun. Mag.*, vol. 40, pp. 108–115, June 2002.

[44] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDMA with adaptive subcarrier, bit and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1747–1758, Oct 1999.

[45] G. Li and H. Liu, "Dynamic resource allocation with finite buffer constraint in broadband OFDMA networks," in *Proc. of the IEEE Wireless Communications and Networking Conference*, Louisiana, USA, March 2003, pp. 1037–1042.

[46] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDMA systems with proportional fairness," *IEEE Trans. Wireless Commun.*, 2005, to be published.

[47] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[48] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, pp. 150–154, Feb 2001.

[49] D. Gesbert, M. Shafi, D. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: An overview of MIMO space-time coded wireless systems," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 281–301, April 2003.

[50] D. Gesbert and J. Akhtar, "Breaking the barriers of shannon's capacity; an overview of MIMO wireless systems," *Telenor Telektronikk*, pp. 53–64, Jan 2002.

[51] W. Ajib and D. Haccoun, "An overview of scheduling algorithms in MIMO-based fourth-generation wireless systems," *IEEE Network*, vol. 19, pp. 43–48, Oct 2005.

[52] J. Wang and B. Daneshrad, "A comparative study of MIMO detection algorithms for wideband spatial multiplexing systems," in *Proc. of the IEEE Wireless Communications and Networking Conference*, Los Angeles, USA, March 2005, pp. 408–413.

[53] Y.-J. Choi and S. Bahk, "Downlink scheduling with fairness and optimal antenna assignment for MIMO cellular systems," in *Proc. of the IEEE Globecom*, Texas, USA, Nov 2004, pp. 3165–3169.

[54] H. Yin and H. Liu, "Performance of space-division multiple-access (SDMA) with scheduling," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 611–618, Oct 2002.

[55] V. K. N. Lau, "Optimal downlink space-time scheduling design with convex utility functions - multiple antenna systems with orthogonal beamforming," *IEEE Trans. Veh. Technol.*, vol. 54, pp. 1322–1333, July 2005.

[56] Y. J. Zhang and K. B. Letaief, "An efficient resource-allocation scheme for spatial multiuser access in MIMO/OFDM systems," *IEEE Trans. Commun.*, vol. 53, pp. 107–116, January 2005.

[57] Y. Cao and V. O. K. Li, "Scheduling algorithms in broadband wireless networks," *Proceeding of the IEEE*, vol. 89, pp. 76–87, Jan 2001.

[58] Y. J. Zhang and k . B. Letaief, "Energy-efficient MAC-PHY resource management with guaranteed QoS in wireless OFDMA networks," in *Proc. of the IEEE International Conference on Communications*, Seoul, Korea, May 2005, pp. 3127–3131.

[59] H. Wang, "Power-sensitive fair scheduling in multiple antenna systems," in *Proc. of the IEEE Vehicular Technology Conference*, Texas, USA, September 2005, pp. 2575–2579.

[60] L. Xu, X. Shen, and J. W. Mark, "Fair resource allocation with guaranteed statistical QoS for multimedia traffic in wideband CDMA cellular network," *IEEE Trans. Mobile Computing*, vol. 4, pp. 166–177, March 2005.

[61] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1150–1158, Nov 2003.

[62] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, pp. 362–370, Dec 2003.

[63] F. Pereira, "A triple user characterization model for video adaptation and quality of experience evaluation," in *International Workshop on Multimedia Signal Processing*, Shangai, China, November 2005.

[64] V. Donini, F. Lironi, C. Masseroni, and R. Trivisonno, "Semantic-aware radio resource scheduling for video streaming in mobile packet networks," in *Proc. IST Mobile and Wireless Communications Summit*, Dresden, Germany, June 2005.

[65] G. Araniti, P. D. Meo, A. Iera, and D. Ursino, "Adaptively controlling the QoS of multimedia wireless applications through user profiling techniques," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 1546–1556, December 2003.

[66] C. E. Luna, L. P. Kondi, and A. K. Katsaggelos, "Maximizing user utility in video streaming applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 141–148, February 2003.

[67] Z. Jiang, Y. Ge, and Y. Li, "Max-utility wireless resource management for best-effort traffic," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 100–111, Jan 2005.

[68] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol. 41, pp. 74–80, October 2003.

[69] G. Carneiro, J. Ruela, and M. Ricardo, "Cross-layer design in 4G wireless terminals," *IEEE Wireless Commun. Mag.*, vol. 48, pp. 7–13, April 2004.

[70] C. K. K. Patabandi, "Multimedia communications over 3G wireless communication systems," Ph.D. dissertation, School of Electronic and Physical Sciences, University of Surrey, 2004.

[71] C. E. Luna, Y. Eisenberg, R. Berry, T. N. Pappas, and A. Katsaggelos, "Joint source coding and data rate adaptation for energy efficient wireless video," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 1710–1720, December 2003.

[72] A. Katsaggelos, F. Zhai, Y. Eisenberg, and R. Berry, "Energy-efficient wireless video coding and delivery," *IEEE Trans. Wireless Commun.*, vol. 12, pp. 24–30, August 2005.

[73] F. Zhai, Y. Eisenberg, and A. K. Katsaggelos, *"Joint source-channel coding for video communications" in Handbook of Image and Video*, 2nd ed. Elsevier Academics Press, 2005, pp. 1–36.

[74] Q. Zhang, W. Zhu, and Y. Zhang, "End-to-end QoS for video delivery over wireless internet," *Proc. IEEE*, vol. 93, pp. 123–134, January 2005.

[75] S. Zhao, Z. Xiong, and Z. Wang, "Optimal resource allocation for wireless video over CDMA networks," *IEEE Trans. Mobile Comput.*, vol. 4, pp. 56–67, January 2005.

[76] Q. Li and M. Schaar, "Providing adaptive QoS to layered video over wireless local area networks through real-time retry limit adaptation," *IEEE Trans. Multimedia*, vol. 6, pp. 278–290, April 2004.

[77] E. Roddolo, G. Panza, C. Lamy-Bergot, P. Amon, M. Martini, G. Jeney, L. Hanzo, and J. Huusko, "Joint source and channel (de)coding in 4G

networks: the phoenix project," in *International Symposium on Wireless Personal Multimedia Communications*, Padova, Italy, September 2004.

[78] M. Livny, B.Melamed, and A. K. Tsiolis, "The impact of autocorrelation on queuing systems," *Management Science*, vol. 39, pp. 322–339, March 1993.

[79] U. K. Sarkar, S. Ramakrishnan, and D. Sarkar, "Modeling full-length video using markov-modulated gamma-based framework," *IEEE/ACM Trans. Networking*, vol. 11, pp. 638–649, August 2003.

[80] H. Zhu, A. Matrawy, and L. Lambadaris, "Models and tools for simulation of video transmission on wireless networks," in *Proc. of IEEE Canadian Conference on Electrical and Computer Engineering*, vol. 2, Ontario, Canada, May 2004, pp. 781–784.

[81] A. Alheraish, S. A. Alshebeili, and T. Alamri, "A GACS modeling approach for MPEG broadcast video," *IEEE Trans. Broadcast.*, vol. 50, pp. 132–141, June 2004.

[82] (2003) MPEG-4 ISO/IEC 14496 video reference software revision fpdam1-1.0-000403. [Online]. Available: http://isotc.iso.ch

[83] M. Ghanbari, *Video Coding: An Introduction to Standard Codecs*, 1st ed. IEE, 1999, ch. 3, pp. 27–32.

[84] E. M. Scheuer and Stoller, "On the generation of normal random vectors," *Technometrics*, vol. 4, pp. 278–281, May 1962.

[85] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, 1992, ch. 2, pp. 96–98.

[86] W. J. Kim, J. W. Yi, and S. D. Kim, "A bit allocation method based on picture activity for still image coding," *IEEE Trans. Image Processing*, vol. 8, pp. 974–977, July 1999.

[87] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye, "Modeling and simulation of self-similar variable bit rate compressed video: A unified approach," in *Proceedings of the ACM SIGCOMM*, vol. 25, Massachusetts, USA, October 1995, pp. 114–125.

[88] (2003) The network simulator version two (ns-2). [Online]. Available: http://www.isi.edu/nsnam/ns/

[89] S. Hara and R. Prasad, *Multicarrier Techniques for 4G Mobile Communications*, 1st ed. Artech House, 2003.

[90] E. Lawrey, "Adaptive techniques for multiuser OFDM," Ph.D. dissertation, School of Engineering, James Cook University, 2001.

[91] Y. Zhang and K. B. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1566–1575, Sept 2004.

[92] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, 1st ed. Cambridge University Press, 2005, ch. 4, p. 176.

[93] *Feasibility Study for Orthorgonal Frequency Division Multiplexing (OFDMA) for UTRAN Enhancement*, 3GPP Std., Rev. TR 25.892 v6.0.0, June 2004.

[94] (2005) it++ v3.8.0 manual. [Online]. Available: http://itpp.sourceforge.net

[95] *Multiplexing and channel coding (FDD)*, 3GPP Std., Rev. TR 25.212 v6.2.0, June 2004.

[96] J. P. Woodard and L. Hanzo, "Comparative study of turbo decoding techniques: an overview," *IEEE Trans. Veh. Technol.*, vol. 49, pp. 2208–2233, Nov 2000.

[97] *Spreading and Modulation (FDD)*, 3GPP Std., Rev. TR 25.213 v3.2.0, December 2004.

[98] *Universal Mobile Telecommunications System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS*, 3GPP Std., Rev. TR 101.112 v3.2.0, April 1998.

[99] J. Javaudin, C. Dubuc, D. Lacroix, and M. Earnshaw, "An OFDM evolution for the UMTS high speed downlink packet access," in *Proc. of the IEEE Vehicular Technology Conference*, Los Angeles, USA, Sept 2004, pp. 846–850.

[100] M. Ghanbari, *Video Coding: An Introduction to Standard Codecs*, 1st ed. IEE, 1999, ch. 7, pp. 173–174.

[101] M. Thelander, "Qualcomm acquires flarion: Bridging the gap between 3G and beyond," *Signals Flash!, Signals Research Group LLC*, pp. 1–4, Aug 2005.

[102] R. Laroia, S. Uppala, and J. Li, "Designing a mobile broadband wireless access network," *IEEE Signal Processing Mag.*, vol. 21, pp. 20–28, Sept 2004.

[103] L. Huang, S. Kumar, and C. C. J. Kuo, "Adaptive resource allocation for multimedia QoS management in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 53, pp. 547–558, March 2004.

[104] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 301–317, March 2001.

[105] J. Matta, C. Pepin, K. Lashkari, and R. Jain, "A source and channel rate adaptation algorithm for amr in voip using the E-model," in *Proceedings of the 13th international workshop on Network and operating systems support for digital audio and video*, New York, USA, 2003, pp. 92–99.

[106] T.-W. A. Lee, S.-H. G. Chan, Q. Zhang, W. W. Zhu, and Y.-Q. Zhang, "Allocation of layer bandwidths and FECs for video multicast over wired and wireless networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, pp. 1059– 1070, Dec 2002.

[107] K. Deb and S. Agrawal, "A niched-penalty approach for constraint handling in genetic algorithms," in *Proc. International Conference on Artificial Neural Networks and Genetic Algorithms*, Portoroz, Slovenia, April 1999, pp. 235–243.

[108] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 2nd ed. John Wiley and Sons, 2001, ch. 14, p. 249.

[109] Z. Diao, D. Shen, and V. Li, "CPLD-PGPS scheduling algorithms in wireless OFDM systems," in *Proc. of the IEEE Global Telecommunications Conference*, Dallas, USA, Nov 2004, pp. 3732–3736.

[110] Y. Zhang and K. B. Letaief, "Adaptive resource allocation and scheduling for multiuser packet-based OFDM networks," in *Proc. of the IEEE International Conference on Communications*, Paris, France, June 2004, pp. 2949–2953.

[111] J. Cai, X. Shen, and J. W. Mark, "Downlink resource management for packet transmission in OFDMA wireless communication systems," *IEEE Trans. Wireless Commun.*, 2005, to be published.

[112] G. Song, Y. Li, L. J. Cimini, and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc. of*

the IEEE Wireless Communications and Networking Conference, Atlanta, USA, March 2004, pp. 1922–1927.

[113] C. Chao, Y. Tseng, and L. Wang, "Dynamic bandwidth allocation for multimedia traffic with rate guarantee and fair access in WCDMA systems," *IEEE Trans. Mobile Comput.*, vol. 4, pp. 420–429, Sept 2005.

[114] A. Stamoulis, N. Sidiropoulos, and G. Giannakis, "Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 512–523, March 2004.

[115] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE*, vol. 93, pp. 57–70, January 2005.

[116] M. Kang and M. Alouini, "Transmission of multiple description codes over wireless channels using channel balancing," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2070–2075, September 2005.

[117] A. Alexiou, D. Avidor, P. Bosch, S. Das, P. Gupta, B. Hochwald, T. E. Klein, J. Ling, A. Lozano, T. Marzetta, S. Mukherjee, S. J. Mullender, C. Papadias, R. Valenzuela, and H. Viswanathan, "Duplexing, resource allocation and inter-cell coordination design recommendations for next-generation wireless systems," *Wireless Communications and Mobile Computing Journal*, vol. 5, pp. 77–93, Feb 2005.

[118] G. Foschini, H. Huang, S. Mullender, S. Venkatesan, and H. Viswanathan, "Physical-layer design for next-generation cellular wireless systems," *Bell Labs Technical Journal*, vol. 10, pp. 157–172, Aug 2005.

[119] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, 1992, ch. 6, pp. 213–219.

[120] A. M. Mood and A. G. Graybill, *Introduction to the Theory of Statistics*, 3rd ed. McGraw-Hill, 1974, ch. 5, pp. 202–203.